Essays on Active Investing

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by
Roger Rüegg
from Weisslingen, ZH

approved in February 2019 at the request of
Prof. Dr. Markus Leippold
Prof. Dr. Michael Wolf

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 13.02.2019

Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgements

Firstly, I want to express my gratitude to Prof. Dr. Markus Leippold, my thesis supervisor, for his excellent guidance, his motivation, and the immense knowledge that I could benefit from. I highly appreciated our insightful discussions and his infectious energy for our research.

I am also very grateful to Prof. Dr. Michael Wolf, my PhD co-examiner, for the time he dedicated to me and the inputs he shared when difficulties arose. His teaching, starting from the very first statistic lecture at university, and research shaped my ideas.

I deeply appreciate all the support I received from my heads at Swisscanto Invest by Zürcher Kantonalbank, Iwan Deplazes, and Dr. René Nicolodi. Without their patience I could have not achieved the research quality.

Furthermore, I had the chance to be in a wonderful team at Zürcher Kanontalbank. My biggest thank goes to Dr. Fabian Ackermann who always helped me to combine my research and work. I would like to thank Florian Arnold, Michael Bretscher, Dr. Andreas Kappler, and Andri Silberschmidt, my colleagues, for always being ready to help and for all the good times I had together with them. Special thanks go to my friends, Martin Gillholm, Meriton Ibraimi, and Istvan Redl, for the discussions and support during the last part of my thesis. Moreover, my studies wouldn't have been as joyful as it had been without my flat mate, Pascal Buri.

Finally, I wish to express my deepest gratitude to my parents, Monika and Martin, and my siblings, Nicole and Kevin, for supporting me in every step toward my PhD. My deepest gratitude goes to my fiancée, Marielle, who encouraged me throughout all the ups and downs.

Winterthur, October 2018 Roger Rueegg

Contents

I Introduction **7**

1 Introduction and Summary of Research Results 9
Roger Rueegg

II Research Papers **13**

1 The Mixed vs the Integrated Approach to Style Investing: Much Ado About Nothing? 15
Markus Leippold and Roger Rueegg

2 Is Active Investing a Zero-Sum Game? 59
Markus Leippold and Roger Rueegg

3 The Long-Only Integrated Approach to Factor Timing 119
Roger Rueegg

III Appendix **167**

1 Curriculum Vitae 169
Roger Rueegg

I

Introduction

Introduction and Summary of Research Results

This dissertation gives answers to three timely questions in the field of active investing. The discussion about the merits of active investing significantly affects the structure of the asset management industry. In addition, the emergence of the so-called style or factor investors that base their investment strategy on the pioneering work of Fama and French (1992) puts further pressure on traditional active management. Thus, we take the opportunity to shed light on important issues that arise in this highly competitive discipline.

The three research papers don't take the recent results in literature for granted. With extensive data samples and a battery of robust statistical tests we highlight different shades of active and factor investing. Because we agree with Bailey et al. (2014) on the fact that the increasing computational power and incentive of institutions to deliver extraordinary results make it crucial to apply the most advanced statistical testing frameworks.

The first research paper, *The Mixed vs the Integrated Approach to Style Investing: Much Ado About Nothing?*, shows that there is no difference in performance between the integrated and the standard mixed approach to style investing. The standard approach regards factors such as bundles of securities and mixes different factor portfolios for the multi-factor investment. On the other hand, the integrated approach regards stocks such as bundles of factors, and invests only in the stocks that share the best factor characteristics on aggregate. Recent literature argues that the integrated approach offers lower risks and higher returns. However, their argumentation contradicts the standard paradigm that higher returns can only be achieved by taking higher risks. We thus build a robust statistical test framework and compare 104 different factor combinations and portfolio constructions during the long history from 1963 to 2016. When we naively test the hypothesis, we arrive at the same conclusion as Bender and Wang (2016), Clarke et al. (2016), or Fitzgibbons et al. (2016). However, we find that the integrated approach by construction has a higher active risk. When we build a fair comparison of the two approaches with similar active risks, the advantage of the integrated approach vanishes, and we can not find statistical evidence for either approach. Still, the integrated approach can offer implementation advantages, as we can see in the third research paper. We also demonstrate

that the integrated approach leads to a higher sensitivity to the low-volatility anomaly.

Our second research paper, *Is Active Investing a Zero-Sum Game?*, explores an extensive dataset of more than 60,000 equity and fixed income mutual funds among different investment categories. For our analysis, we build a novel statistical framework that takes the observed cross-sectional and serial dependence of the mutual funds' returns into account. At the same time, it adjusts for the multiple hypothesis problems that arise with different fund providers and investment categories. Our results show that we cannot reject the hypothesis of a zero-sum game between active and index investing for a vast majority of investment categories. Thus, we find evidence for the theory of Berk and Green (2004), who demonstrated that rational markets lead to a zero-sum game after fees. When we analyze the performance drivers of the difference between active and index investing, we expected active management to protect investors from sudden volatility shocks. Counter-intuitively, we find that active management tends to outperform in calm market environments and to be negatively affected during crisis periods. We also find that active equity relative to index funds show a positive exposure to small-cap and growth stocks while active fixed income relative to index funds show a higher sensitivity to credit risk. Contrary to that, index managers exhibit a higher sensitivity to the traditional market and duration risk premium. When we investigate the role of performance persistence, fees, and size, we find that active low-fee winner portfolios and active small winner portfolios tend to outperform index investors. However, their alpha does not survive our robust test statistics. On the other hand, our results show significant negative alphas as well after the multiple hypothesis adjustment for active equity retail investors that invested in high-fee losers.

In the third research paper, *The Long-Only Integrated Approach to Factor Timing*, I try to time the factors in a realistic long-only setting. It shows that a Markov switching model with one month lag and two states can generate an alpha of 0.36% per month. The alpha is adjusted for the underlying factor exposures and thus reflects the timing contribution of the strategy. Hence, this adds evidence to the recent findings of factor momentum by Arnott et al. (2018). In contrast to their long-short mixed approach, I show that factor momentum also works in the highly transparent long-only integrated approach. Moreover, the timing ability exists not only in the US but also in the developed and emerging markets, among different factor sets, and for holding periods of up to 12 months. Most of the combinations tested survive the robust alpha test that we developed in the second research paper.

One caveat of short-term timing strategies is the high turnover. Trading costs may erase the gains in markets with high transaction costs. When I reduce rebalancing frequencies in markets with high transaction costs to limit the turnover, the alphas stay positive and mostly survive the robust test statistics. However, the significance vanishes when I adjust for multiple hypothesis. Still, I achieve an alpha of 0.29% per month after transactions costs which looks economically significant. Hence, there is evidence that the Markov switching strategy may offer a promising source of alpha, which implies that the market prices of risk adjust only slowly over time.

References

Arnott, Robert D., Mark Clements, Vitali Kalesnik, and Juhani T. Linnainmaa, 2018, Factor momentum, Available at SSRN 3116974.

Bailey, David H., Jonathan M. Borwein, Marcos L. de Prado, and Qiji J. Zhu, 2014, Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance, *Notices of the American Mathematical Society* 61, 458–471.

Bender, Jennifer, and Taie Wang, 2016, Can the whole be more than the sum of the parts? Bottom-up versus top-down multifactor portfolio construction, *Journal of Portfolio Management* 42, 39–50.

Berk, Jonathan B., and Richard C. Green, 2004, Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269–1295.

Clarke, Roger G., Harindra De Silva, and Steven Thorley, 2016, Fundamentals of efficient factor investing, *Financial Analysts Journal* 72, 9–26.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

Fitzgibbons, Shaun, Jacques Friedman, Lukasz Pomorski, and Laura Serban, 2016, Long-only style investing: Don't just mix, integrate, Integrate (June 29, 2016), AQR Capital Management, LLC.

II

Research Papers

The Mixed vs the Integrated Approach to Style Investing: Much Ado About Nothing?

Markus Leippold and Roger Rueegg

This paper is published in:

Leippold, Markus, and Roger Rueegg, 2018, The mixed vs the integrated approach to style investing: Much ado about nothing?, *European Financial Management* 24, 829–855.

DOI: https//:doi.org/10.1111/eufm.12139

I have presented it at:

- UZH Brown Bag Doctoral Lunch Seminar, February 2017, Zürich, Switzerland.
- Citigroup's 14th Quantitative Research Conference, June 2017, Budapest, Hungary.
- EFMA's 26th Annual Meeting, June 2017, Athens, Greece.

Abstract

We study the difference between the returns to the integrated approach to style investing and those to the mixed approach. Unlike the mixed approach, the integrated approach aggregates factor characteristics at security level. Recent literature finds that the integrated approach dominates the mixed approach. Using statistical tools for robust performance testing, we demystify these findings as a statistical fluke. We do not find any evidence favoring the integrated approach. What we do find is that the integrated approach exhibits a higher sensitivity to the low-risk anomaly. However, this reduction in risk does not lead to an improvement in performance.

1 Introduction

Style investing is the investment process that aims to harvest risk premia through exposure to factors. Factors are the foundation of all portfolios: they are the persistent forces driving the returns of stocks, bonds, and other assets. There are diverging views on how to build multi-factor portfolios. The current debate is centered around two approaches. The first approach is to mix where a portfolio is built by combining stand-alone factor portfolios. The second approach is to integrate where a portfolio is built by selecting securities that have simultaneously strong exposure to multiple factors at once. Recent research suggests that a bottom–up or integrated approach provides higher returns and lower risks than a mixed approach.[1] Hence, it seems that the debate about mixing or integrating has been concluded.

However, such a finding clearly must invite suspicion, as it contradicts the standard paradigm in finance. Higher returns can only be achieved by taking higher risks.[2] We contribute to the recent literature on style investing by providing a thorough analysis of the differences in the returns and risk of the mixed and integrated approaches to long-only style investing. We find that the integrated approach shows superior returns to risk characteristics in only a few combinations of styles. When we adjust for multiple hypothesis testing, we can no longer reject the hypothesis that the two approaches are the same. Hence, our findings present a challenge to the previous literature that promotes the integrated approach.

An early contribution that analyzes an integrated (or bottom–up) approach to style investing is Haugen and Baker (1996). With a selected set of factors, they show that factor models are surprisingly accurate in forecasting the future relative returns of stocks. They find high abnormal returns together with lower risk numbers in stocks with high predicted returns and argue that their result reveals a major failure in the efficient markets hypothesis. Subsequent contributions on style investing turned their focus on how to optimally combine individual factor portfolios. Of interest was not whether to

[1] See Bender and Wang (2016), Clarke et al. (2016), and Fitzgibbons et al. (2016).

[2] In addition, it contradicts the risk based explanation of why style premia exist. To clarify this point, consider a combination of value and momentum stocks. Strictly speaking, we thereby avoid overvalued momentum stocks that are threatened by a sudden market crash. However, bearing the risk of a sudden crash is the most rational explanation of why the momentum premium exists (see Daniel and Moskowitz (2015)). Hence, we would expect lower returns in the integrated approach, contrary to what recent publications suggest.

mix or integrate, but on how to derive optimal factor exposures.[3] Factor portfolios were regarded as given building blocks.

Only recently, the integrated approach regained attention with two recent publications. Clarke et al. (2016) argue that the mixed approach captures only one-half of the potential improvement over the market Sharpe ratio. They show that when the group constraint is released and the securities are viewed as a bundle of styles instead of the styles being regarded as a bundle of securities, one can capture much more of the excess returns of the factors. The second work promoting the integrated approach is Bender and Wang (2016). They assert that integration leads to a superior risk–return trade-off due to the fact that it captures nonlinear cross-sectional interaction effects between factors.

Interestingly, the ETF industry has yet to make up its mind whether mixing or integrating is the right approach. Table 1 summarizes the most well-known multi-factor ETFs. While the largest ETF, managed by Goldman Sachs, pursues a mixed approach to factor investing, we find FlexShares, JP Morgan, and iShares implementing an integrated approach to factor investing. Moreover, AQR, one of the largest global investment managers with $159.2 billion assets under management as of August 2016, maintain in Fitzgibbons et al. (2016) that a long-only portfolio is more profitable if based on an integrated approach. Indeed, it is the general tenet of the financial industry that the integrated approach is superior to the mixed approach. To our best knowledge, the only contrarian view that we are aware of is the white paper by Fraser-Jenkins et al. (2016). They find that the integrated and mixed approaches lie on the same return to risk line.

[Table 1 about here.]

Given the inconclusive evidence, we conduct an in-depth analysis of the two long-only methodologies of style investing and contribute to the literature in several ways. First, we analyze all the combinations of the Fama and French (2015) five-factor model extended by the momentum and low volatility factors. Moreover, we analyze an extended period from 1963 to 2016 of all NYSE, AMEX, NASDAQ stocks. By doing so, we expand on the previous literature that concentrates on only a few combinations of styles and markets, mostly on a shorter time period. For example, Fitzgibbons et al. (2016) analyze the combination of value and momentum from 1993 to December 2015, and Bender

[3] See, e.g. Blitz (2015) for an overview.

and Wang (2016) the six possible two- and four-factor combinations of value, low volatility, quality, and momentum, from 1993 to March 2015. Clarke et al. (2016) analyze the four-factor combination of low beta, size, value, and momentum, from 1968 to 2014.

Second, and more importantly, we extend the comparison of the two approaches by building a robust multiple hypothesis framework. While the previous literature reports simple risk and return differences and finds economically sound advantages to the bottom–up construction, we question these results. Motivated by Bailey et al. (2014), who argue that shallow statistical analysis can easily lead to allocating capital to strategies that were false discoveries, we apply a set of robust performance tests to the hypothesis that the integrated approach offers a better performance than the mixed approach. To avoid backtest overfitting, we adjust for the number of portfolio combinations tried. Such an adjustment is common in medical research. Yet, in finance, multiple hypotheses methods have only recently gained attention.[4] Hence, we hope that our study will increase the awareness that the lack of rigorous statistical procedures might lead to wrong and misleading conclusions. Other papers that have recognized the importance of multiple hypothesis testing include Leippold and Lohre (2012a,b, 2014).

Our empirical results are as follows. When we follow the argumentation of the previous literature, we can confirm their results in that the integrated approach is superior. However, when we apply our battery of more robust statistical tests and include all possible style combinations as well as a longer time horizon, we must conclude that the performance differences are statistically insignificant. What we also find is that the risk of the integrated approach may be lower than that of the mixed approach. Furthermore, it turns out that the integrated approach shows a high sensitivity to the low-risk anomaly, originally discovered by Jensen et al. (1972). This result confirms our intuition behind the integrated approach, which is one of avoiding risk through broader diversification. However, we also find that the lower risk is accompanied by lower returns. Hence, the risk reduction of the integrated approach does not lead to an improvement of performance. When we further analyze trading costs and turnover, we find on average a lower turnover in the integrated approach. This can lead to significant differences in selected portfolio construction techniques and style combinations when trading costs are high. However, due to the low trading costs nowadays we observe no significant

[4] See, e.g., Harvey et al. (2016) for a current discussion.

difference in the most recent past.

The paper proceeds as follows. In Section 2, we present the methodology behind our portfolio construction and hypothesis testing. Section 3 presents our data and factor choice. In Sections 4, we summarize our empirical findings. In Section 5, we provide a turnover analysis and point at possible limitations. Section 6 concludes.

2 Methodology

We now provide some guidance on the methodology for constructing portfolios in the mixed and integrated approach. Then, we briefly present the statistical framework for testing whether the integrated approach to style investing offers higher risk-adjusted excess returns than the mixed approach.

2.1 Portfolio construction

We assume that we have $i = 1, ..., n$ securities and $f = 1, ..., k$ styles with the style information matrix $\Phi \in \mathbb{R}^{n,k}$. Each column $\phi_f \in \mathbb{R}^n$ of Φ contains the style figures for the n securities. For example, for the style 'value' these figures are the book-to-market ratios of the companies. Each security obtains for each factor a score $s_{i,f}$ based on the style information. There are two common ways to build this score, the rank-based and z-score approach. The rank-based score neglects the distribution of ϕ_f and scores the securities among their ranks. We build the score as

$$s_{i,f}^{\text{rank}}(\phi_f) = \frac{\text{rank}(\phi_{f,i}, \phi_f) - 1}{n - 1}, \tag{1}$$

where the operator 'rank' runs from 1 to n from the smallest to the largest values in ϕ_f. The score is invariant to the numbers of securities and lies between 0 (worst) and 1 (best). On the other hand, the distribution of ϕ_f is taken into account in the z-score approach. Here, the score is defined as

$$s_{i,f}^{z} = \frac{\phi_{f,i} - \mu(\phi_f)}{\sigma(\phi_f)}. \tag{2}$$

In the mixed approach, we express the single style portfolios $w_f \in \mathbb{R}^n$ as a function φ of the score vector s_f^j,

$$w_f = \varphi(s_f^j), \tag{3}$$

where $j = \{\text{rank}, z\}$. These portfolio weights are then aggregated to the final weights of the mixed approach by giving a weight of a_f to each style portfolio:

$$w_{\text{mix}} = \sum_{f=1}^{F} a_f w_f. \tag{4}$$

In the integrated approach, the aggregation of the style information occurs before constructing the portfolio. For this purpose, we build an aggregated score as follows:

$$s_{\text{agg}}^j = \sum_{f=1}^{F} a_f s_f^j, \tag{5}$$

where we set the weight a_f of each score equal to the style factor portfolios' weight of the mixed approach.[5] To build the integrated portfolio, the same portfolio construction function φ is applied as for the single style portfolios, but the input score vector is the aggregated score s_{agg}^j:

$$w_{\text{int}} = \varphi(s_{\text{agg}}^j). \tag{6}$$

The main difference between the mixed and integrated portfolios is that the mixed portfolio starts with style portfolios based on single scores and then aggregates the information by mixing the style portfolios. In the integrated approach, the information aggregation occurs before constructing the portfolio. Basically, we are free to choose the portfolio construction function φ and score methodology. While for the score methodology, we restrict ourselves to the rank and z-score, we apply four different portfolios methodologies. The first two are analogous to Fama and French (1992) (TER and DEC) and, for the benchmark-sensitive investors, we additionally include the portfolio construction techniques of Bender and Wang (2016) (BW) and Fitzgibbons et al. (2016) (TE) into our analysis.

[5] Concerning the choice of the style portfolio and score weights a_f, we follow DeMiguel et al. (2009) and apply the most naive diversification rule $a_f = \frac{1}{F}$.

Hence, we end of with the following set of portfolio construction methodologies,

$$\mathcal{P} = \{\text{TER},\ \text{DEC},\ \text{BW},\ \text{TE}\}, \tag{7}$$

which we briefly discuss next.

2.1.1 Tercile and decile portfolios: TER and DEC

The tercile (TER) and decile (DEC) portfolio construction follows closely the style portfolio construction originally suggested by Fama and French (1992). First, we build the scores with the rank-based methodology in Equation (1). Second, the function φ of Equations (4) and (6) is such that we invest value-weighted in the upper tercile of the scored companies for the TER and in the upper deciles of the scored companies for the DEC approach.

To clarify the differences between the mixed and integrated approaches, we provide a stylized example for the TER portfolios. We assume that there are 10 stocks, from stock A to stock J, with a given market capitalization (mc) and two factors $f = \{V, W\}$, say, 'value' (V) and 'momentum' (M). Exact numbers are shown in Table 2. The three highest ϕ_V and ϕ_M figures are highlighted in bold. The three highest book-to-market ratios are 0.51 for stock A, 0.82 for stock D, and 0.97 for stock I. The highest returns for the past 12 months disregarding the most recent month are 0.09 for stock A, 0.14 for stock B, and 0.22 for stock I. We first build the 'value' and 'momentum' scores s_V^{rank} and s_M^{rank} as illustrated in Equation (1) and take their the average to arrive at the aggregated score $s_{\text{agg}}^{\text{rank}}$ shown in Equation (5).

[Table 2 about here.]

For the mixed portfolio, we first apply the portfolio construction function φ to the scores s_f^{rank} and value-weight the upper tercile of the stocks. This procedure results in the single style portfolios w_V and w_W. The final weights are simply the average of the factor portfolio weights and are shown in column w_{mix}. In contrast, the integrated approach aggregates the information on security level. We first build the aggregated score $s_{\text{agg}}^{\text{rank}}$ that is the average of the style scores s_V^{rank} and s_M^{rank}. Given the aggregated score, we then value-weight the upper tercile of the stocks as shown in column w_{int}.

We end up with four groups of stocks. The first group is assigned a weight of zero in either approach. They show neither superior style characteristics, nor are they on average superior to the other stocks. The second group of stocks is only represented in the mixed portfolio, but not in the integrated approach. They show superior style characteristics in one of the two styles. However, the score in the second style is too small to be considered in the integrated portfolio. The third group of stocks is not considered in any of the style portfolios, but it shows superior style characteristics when all styles are aggregated. Stock C is an example of this group. The fourth group of stocks belongs in both the mixed and integrated portfolios.

The mixed portfolio always holds at least as many stocks as the integrated portfolio. The more similar the styles, the fewer stocks are included in the mixed portfolio. If each style provides the exact same ranking of stocks, the weights of the integrated and mixed approaches are equal. With two (three) styles and without an overlap, it is possible to hold 60 (90) percent of the stocks in the mixed portfolio. For more than four styles, the mixed portfolio could possibly hold all the stocks of the universe, while on the other hand the integrated approach holds by definition in any case 30 percent of the stocks in the universe.

2.1.2 Bender and Wang (2016) portfolios: BW

In contrast to the tercile and decile portfolios, the portfolio construction of Bender and Wang (2016) concentrates on under- and overweighting relative to the market-cap-weighted benchmark. The scores are built on the z-score methodology in Equation (2). The portfolio construction function φ is defined by first ordering the securities according to the score and grouping them into 20 subportfolio with each holding 5 percent of the total market capitalization. For each group, a multiplier of 0.05 to 1.95 with increments of 0.1 is applied to the market-cap weight of the securities. For example, the subportfolio of companies with the highest (lowest) score gets its market-cap weight multiplied by 1.95 (0.05). In the last step, the weights are normalized. This procedure results in an over- and underweighting of the highest and lowest scored companies. For a detailed description of the resulting portfolios, we refer to the original work of Bender and Wang (2016). Compared to the TEC and DEC portfolios of the previous section, the BW portfolios invest in all securities in the mixed and integrated approach.

2.1.3 Target tracking error portfolios: TE

We additionally implement the target tracking error optimization suggested in Fitzgibbons et al. (2016). The portfolio construction function φ is defined as

$$\varphi(s) := \max_w (w - w_M)'s \quad \text{s. t.} \quad \sqrt{(w - w_M)'\Sigma(w - w_M)} \leq \sigma_{te}, \tag{8}$$

where we construct the score s with the z-score methodology in Equation (2) for the single style portfolios of the mixed approach and with the aggregated score in Equation (5) for the integrated approach. Furthermore, by σ_{te} we denote the ex-ante target tracking error, by w_M the market-cap-weighted benchmark, and by Σ the covariance matrix of the returns. Since we deal with a large covariance matrix, we use the shrinkage methodology of Ledoit and Wolf (2004) based on the most recent 24 observations. To obtain similar levels of tracking errors as in the BW approach, we set the ex-ante target tracking error σ_{te} to two percent annualized.

2.2 Multiple hypothesis testing

As Bailey et al. (2014) argue, researchers and financial institutions are incentivized to try several possibilities, but report only the significant results. In our case, we have 5 styles that result in 26 possible combinations. Therefore, our hypothesis is split into 26 individual hypotheses. To make the case for multiple hypothesis testing, we provide a simple illustration with a momentum-based strategy. To this end, we consider IBM stock with a history of returns data from January 1960 to December 2014. We create 20 momentum strategies. The first strategy invests in IBM for the next month if the previous month was positive, otherwise it steps out of the market. The second strategy analogously invests in IBM for the next month if the second most recent month was positive, and disinvests if it was negative. This analysis is conducted for the most recent 20 months, which results in 20 different momentum strategies. Our main goal is to test whether a specific momentum strategy shows a significantly higher Sharpe ratio than the buy-and-hold strategy. For the test statistics, we consider the robust Sharpe ratio test of Ledoit and Wolf (2008).

When testing 20 different momentum strategies as specified above, we have to be sensitive to

the fact that the probability of finding no significant results at a confidence level of 5 percent is $(1-0.1)^{20}$ or approximately 12 percent. Therefore, the complementary probability of finding at least one significant momentum strategy by pure luck is roughly 88 percent. By inspection of Table 3, we find that there are three Sharpe ratios significantly different from the buy-and-hold strategy, two significantly lower and one significantly higher. In particular, the return of the momentum strategy with the 16th look-back month shows an annualized return of 7.1 percent with a annual volatility of 16.5 percent. For comparison, the annualized return of IBM over the period June 1963 to December 2014 was lower, at 4 percent, with a higher annualized volatility of 23.8 percent. We can reject the hypothesis that the Sharpe ratio is equal to the buy-and-hold strategy at a confidence level of 10 percent. However, recalling that we tested 20 different strategies and to be sure that our superior strategy is not just a statical fluke, we must embed our p-values into a robust multiple hypothesis framework.

[Table 3 about here.]

There are many ways to deal with the problem of multiple hypothesis testing.[6] For our analysis, we focus on the family wise error rate (FWER), the probability of at least one false discovery. The FWER is the most common approach to control for multiple hypothesis. For a large number of hypothesis the FDR was introduced in Benjamini and Hochberg (1995) and is defined as the expectation of the proportion of falsely rejected null hypotheses. But Romano and Wolf (2005a) and Romano and Wolf (2005b) point out that they make the strong assumption that the individual p-values are independent of each other. Therefore, they propose a resampling-based stepdown multiple testing framework that considers the dependence structure of the test statistics. Their method comes at a high computational cost. In Romano and Wolf (2016), the authors refine their method. Not only are they able to reduce the computational cost, but their method also avoids choosing a fixed significance level α. Hence, their framework allows for dependence structures in the test statistics without loss in statistical power.

For comparison and as a robustness test, we also focus on the older tests of Bonferroni (1936) and Holm (1979). The Bonferroni test divides the required significance value by the number of hypotheses. A confidence level of 5 percent with 20 tries produces a threshold of $0.05/20 = 0.0025$. As pointed out

[6] For a comprehensive overview, we refer to Harvey et al. (2016).

by Conneely and Boehnke (2007), the Bonferroni adjustment is too conservative for correlated tests. Holm (1979) developed a sequential Bonferroni method, preserving its flexibility but increasing its power. Strictly speaking, the test applies the Bonferroni adjustment only on the subset of hypotheses that are not rejected from the beginning.[7]

How does our previous conclusion from the analysis of the momentum strategy change if we adjust the single hypothesis p-values for multiple hypothesis testing? From Table 3, we observe three significant p-values when we test naively; however, with the multiple hypothesis adjustment, the p-values are all insignificant. Therefore, we can not reject the hypothesis that one of the momentum strategies is different from a simple buy-and-hold strategy. Hence, their abnormal performance is probably a false discovery and nothing but a statistical fluke.

3 Data

We now present our data and factor choices for the multiple hypothesis testing.

3.1 US data from 1963 to 2016

To construct the integrated and mixed portfolios, we use stock return and balance sheet data from the merged CRSP and Compustat database. The stock universe consists of all NYSE, AMEX, and NASDAQ stocks with share codes 10 or 11. The data items of the CRSP and Compustat database are merged by the eight-character CUSIP. We exclude finance, insurance, and real estate companies with SIC codes between 6000 and 6799. We limit the universe to big stocks as defined by Fama and French (1992). This universe consists of 810 stocks on average, starting with the 583 largest US companies in June 1963 and ending with the largest 867 companies in December 2016. At its peak in February 2000, there were 1,548 companies in the universe. The market capitalization breakpoint, which splits the universe into small and large caps, builds the median of all NYSE stocks. The average market capitalization over the analyzed period was US $792 million. It reached a minimum of US $63 million in December 1974 and peaked in June 2014 with US $2,871 million. The universe represents the

[7] Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) show that under dependency, it is favorable to control for the FDR, and not for the FWER.

largest companies in the most important equity market of that time. Consequently, it is a highly liquid universe, tradeable with small transaction costs.

The CRSP database has a monthly frequency and starts in January 1960. As outlined by Fama and French (1992), the data of the Compustat database is not reliable before 1962. Therefore, we use balance sheet information at a yearly frequency from the Compustat database starting at the earliest possible year, in 1962. Moreover, all data points of the Compustat database are lagged by 6 months to guarantee that the balance sheet data of the companies are available at the date of the portfolio construction. Considering the lag of 6 months for the balanced sheet data, the backtest period consists of 63.5 years. It starts in June 1963 and ends in December 2016.

3.2 Factor choices and factor combinations

The most prominent factor model is Fama and French (1992) with the three factors market excess return (M), 'size' (S), and 'value' (V). Carhart (1997) extended the Fama and French factors with the 'momentum' factor (W) of Jegadeesh and Titman (1993). The recent literature has put a lot of effort into detecting other factors that show high abnormal returns. Harvey et al. (2016) find 315 published factors with ostensibly significant excess returns. Among the most recent findings are the 'quality' premium as defined by profitability, growth, safety, or payout prevail anomalies.[8] Consequently, Fama and French (2015) extended their three-factor model to include certain quality aspects with the 'profitability' factor of Novy-Marx (2013) and the 'investment' factor of Aharoni et al. (2013). They argue that expected returns are not solely driven by the book to market ratio (V), but also by 'profitability' (R), and 'investment' (C). Another highly popular anomaly is the 'low-risk' (L) anomaly, originally discovered by Jensen et al. (1972). Empirically, low-beta stocks exhibit higher returns than implied by their market beta. Among many other low-risk measures, Ang et al. (2006) find that stocks with high idiosyncratic risk earn abnormally low average returns.[9]

Without loss of generality, we focus on the most important and widespread style factors: 'value', 'profitability', 'investment', 'momentum', and 'low volatility'. We argue that if the integrated ap-

[8] A good overview of the quality factor is given by Asness et al. (2014).

[9] Recent overviews on the low-risk anomaly include, e.g., Blitz and Van Vliet (2007) and Baker et al. (2011). As a result of the high risk-adjusted returns of the low-risk stocks, every index or smart-beta provider offers a low-volatility product. Data compiled by Bloomberg show that the 10 largest low-volatility or minimum volatility ETFs held $40 billion in assets as of mid-2016.

proach obtains higher risk-adjusted returns compared to the mixed approach, higher risk-adjusted returns should also be observed for the combinations of these most popular factors. Moreover, we concentrate on independent risk factors to arrive at a meaningful analysis. Hence, our analysis is built on the following set of factors $\mathcal{F}$:

$$\mathcal{F} = \{V, R, C, W, L\}. \tag{9}$$

We measure V by book equity as defined in Fama and French (1992) divided by the market capitalization of the CRSP database. We lag book equity by 6 months in order to guarantee that the balance sheet data is published at the date of the portfolio construction. In contrast, the market capitalization is not lagged. Asness and Frazzini (2013) show that this small detail is superior in terms of performance and when the portfolio is rebalanced monthly. The factors R, C, and W are defined as in Fama and French (2015). R is calculated by the ratio of operating profitability divided by book equity and C by the total book assets of the recent year divided by the actual total book assets. All these variables are calculated with data from the Compustat database and lagged by 6 months. W is calculated as the total returns over the past 12 months, while the most recent month is ignored. For L we take the volatility to be the standard deviation of the most recent 36 monthly returns. The look-back period of 36 months is chosen analogously to Blitz and Van Vliet (2007). We show the summary statistics of the factors including the size factor in Table 4.

[Table 4 about here.]

It would be beyond the scope of this paper to test all possible available factors. Instead, we want to give a comprehensive overview of the most popular factors. Moreover the factors should also be independent of each other. For example, Fama and French (1992) also test the earnings-to-price ratio. Since this ratio is highly correlated with the book-to-price ratio, we do not include it in our study, to avoid potential problems arising from multicollinearity. Therefore, before we proceed, we test our selected factors for multicollinearity by calculating the variance inflation factor (VIF). Table 5 reports the results. We find that the VIF stays below two, far below the threshold of 10 which, according to O'Brien (2007), is equivalent to a confidence level of 0.1. Therefore, there is no sign of a linear

dependency in our factor selection (9).

[Table 5 about here.]

Given the 5 following styles, 'value' (V), 'profitability' (R), 'investment' (C), 'momentum' (W), and 'low-volatility' (L), it is possible to build 10 combinations with 2 factors, 10 combinations with 3 factors, 5 combinations with 4 factors, and 1 combination with 5 factors. In all, we end up with 26 possible combinations, denoted by $\mathcal{C}$,

$$\begin{aligned}\mathcal{C} = \{ & VW, VC, VR, VL, WC, WR, WL, CR, CL, RL, \\ & VWC, VWR, VWL, VCR, VCL, VRL, WCR, WCL, WRL, CRL, \\ & VWCR, VWCL, VWRL, VCRL, WCRL, VWCRL\},\end{aligned} \tag{10}$$

where we indicate each combination by the acronym formed from its factors' names.[10]

4 Robust hypothesis testing

We now run the integrated and mixed portfolios for all 26 combinations of $\mathcal{C}$ and all four portfolio construction methodologies in $\mathcal{P}$ in (7). The portfolios are rebalanced monthly. We first compare the Sharpe ratio of the resulting 104 strategies and then concentrate on the variance and the benchmark orientated figures, namely the information ratio and the tracking error.

4.1 Multiple hypothesis testing for the Sharpe ratio

In Figure 1, we report the Sharpe ratios of all 104 strategies in grey and highlight the differences of the integrated to the mixed approach in green (positive) or red (negative). The integrated approach obtains a higher Sharpe ratios in most of the cases. For the BW portfolio construction methodology, we find the integrated approach to outperform the mixed approach in any of the style combinations.

[Figure 1 about here.]

[10] For example, the combination of 'value' (V), 'momentum' (M) and 'low-volatility' (L), is called VWL.

Bender and Wang (2016) find the highest difference in risk-adjusted returns for the combination of 'value', 'low volatility', 'quality', and 'momentum', with 0.84 in the integrated approach and 0.73 in the mixed approach. We too find the highest Sharpe ratio for the 4-factor combinations. For example *VWCR* or *VWRL* show, from June 1963 to December 2016, very high Sharpe ratios: 0.48 and 0.49. In contrast, the mixed approach obtains a Sharpe ratio of 0.41 for both combinations. When we regard the style 'robust' as a proxy for 'quality', we arrive for the combination *VWRL* to the same magnitude of improvement as Bender and Wang (2016). Fitzgibbons et al. (2016) find increasing benefits with the number of uncorrelated factors combined in the portfolio's construction for the post-1993 period. We can also find support for this observation. We find only positive differences for combinations with more than three styles, while we observe only small improvements for the two-factor combinations.

Next, we perform the single hypothesis robust Sharpe ratio test of Ledoit and Wolf (2008). The robust Sharpe ratio test requires an optimal block size for the dependent block bootstrap.[11] Since we observe the highest autocorrelation for the block sizes five, we use this value for the optimal block size (bl).[12] Figure 2 shows the monthly Sharpe ratio differences of the four portfolio construction methodologies and the 26 factor combinations in bars as well as their p-values in symbols. The significant differences at the 95 percent confidence level are highlighted in green (positive) or red (negative). For the TER (DEC) construction, we observe only one (three) significant single hypothesis tests. For the benchmark-orientated portfolio construction method BW (TE), we find 16 (15) p-values to be below the five percent level. Hence, at least under a single-hypothesis test, there seems to be some pattern emerging in favor of the integrated approach when applying the benchmark-oriented construction methods.

[Figure 2 about here.]

However, since we deal with four portfolio construction methods and 26 factor combinations, it is crucial to adjust the single hypothesis p-values in Figure 2 for the numbers of tries. For example, Clarke et al. (2016) focus on the ex-ante portfolio construction and the comparison of the two ap-

[11] There are methods to evaluate the optimal block length for a single time series, see Politis and White (2004) and Patton et al. (2009), or in the bivariate case, see Ledoit and Wolf (2008). But for the multivariate case with more than two time series, there is no framework available.

[12] For robustness we also applied a block size of two, that shows the second highest autocorrelation. The different block size has no impact in all the results presented below and lead to the same conclusions.

proaches. They make the – in their words – 'implausible' assumption that the investor first has the knowledge of the successful 4 styles as well as their information ratio. We argue that exactly this implausible assumption is the game changer. First, the expected information ratios are unknown ex ante. It may well happen that the successful styles of the past will fail in the future. Second, the investor is not aware which styles or which combination will be successful in the future. Therefore, the multiple hypothesis framework that we apply next is of crucial importance. We expect the different portfolio construction methods for the same style combination to behave similarly. Since the framework of Romano and Wolf (2016) accounts for these inherent dependence structures in the test statistics, it is best suited to adjust the single hypothesis p-values.

[Figure 3 about here.]

Figure 3 provides the adjusted p-values (in symbols) under the multiple hypothesis testing framework. We find that all the significant strategies of the TER and DEC portfolio construction are likely to be false discoveries. However, for the benchmark-sensitive investors, we still find five combinations including VW in the framework of BW and five combinations including WL in the TE framework to survive the adjustment and to perform significantly better.

4.2 Variance, information ratio, and tracking error comparison

We now test the null hypothesis that the variance, the information ratio, and the tracking error of the integrated and mixed approach are equal for the 26 factor combinations and four portfolio construction methodologies. Figure 4 reports the differences in the logarithmic variance (top chart), the information ratios (middle chart), and the logarithmic tracking error (bottom chart) based on monthly returns using the integrated and mixed approach. For the variance and tracking error hypothesis testing, we apply the robust variance test of Ledoit and Wolf (2011). Analogous to the previous section, we chose the block size of five for the dependent block bootstrap that is also required for the robust variance test. The only difference to the robust hypothesis testing with the Sharpe ratio is that the analyzed returns are the excess returns above the one-month Treasury bill rate (for the information ratio) and the excess returns above the market-cap weighted benchmark (for the variance and tracking error).

[Figure 4 about here.]

We find eleven combinations for the TER, three for the DEC, 15 for the BW, and one for the TE approach that exhibit a lower variance in the integrated approach, while only the combination *VR* shows a significantly higher variance in the DEC portfolio construction methodology. It is not surprising that the optimized portfolio construction method TE results in similar risk levels, since including the information of the covariance matrix during the portfolio construction leads to neutralized bets in the risk dimension. The more factors included in the strategies, the higher is the percentage of significant differences where we can reject the null hypothesis.

For the portfolio construction methodologies TER and DEC, we observe in general a lower information ratio compared to the mixed approach. The combination *VWL* in the DEC and *VWCL*, *VWRL* as well as *VWCRL* in the TER are significantly lower compared to the mixed approach. On the other hand, we observe in the benchmark-orientated approaches BW and TE on average an improvement in the information ratio. However, none is significantly different from zero. Again, when we take 'robust' as a proxy for quality, we can confirm the result of Bender and Wang (2016) that the integrated approach shows a higher information ratio over time. However, these differences not significantly different from zero, with adjusted p-values close to one.

For the tracking error, we first observe that the integrated approach obtains a significantly higher tracking error in all of the combinations. Second, the differences increase with the numbers of factors. This result comes not as a surprise, since we construct the single style portfolios of the mixed portfolio in the same way as the final portfolio of the integrated approach. Consequently, the equal-weighted average of the single style portfolios in the mixed approach shows a significant difference in the active risk of the two portfolios. This finding is due to the high diversification effect of the single style portfolios that exhibit a low correlation among each other.

4.3 Similar active risk: a fair comparison

So far, we do find some support, although weak, for the results presented in recent literature, in that the integrated approach shows a significantly higher Sharpe ratio for some of the style combinations. However, we also find significantly higher tracking errors compared to the mixed approach and no

improvement in the information ratio. Hence, a fair comparison between the integrated and mixed approach demands some further investigation. To this end, we construct the portfolios such that the same level of active risk is achieved in both methodologies.[13]

4.3.1 Portfolio construction

To increase active risk in the mixed approach, we change the portfolio construction function from Equation (3). Instead of investing in the 30 percent best stocks for the single mixed TER style portfolios, we invest market-cap-weighted in the 20 (two factors), 12.25 (three factors), 10 (four factors) and 8.75 (five factors) percent best stocks. We expect that this increased concentration results in higher tracking errors for the mixed portfolios. The DEC portfolio construction approach is neglected, since it already has a high level of concentration in the integrated approach that is hard to generate in the mixed approach. For the BW portfolio construction method, we take the multiplier ranging from 0.05 to 1.95 to the power of two (two factors), three (three factors), five (four factors), and eight (five factors) for the single style portfolios w_f. By doing so, we give a higher overweight to the five percent market-cap groups with a high score and increase the underweight of stocks with a small score, relative to the market-cap-weighted benchmark. Analogous to Fitzgibbons et al. (2016), we increase the ex-ante annual tracking error target to 3.0 (two factors), 3.5 (three factors), 3.8 (four factors), and 4.0 (five factors) in the TE mixed portfolios. The portfolio construction of the integrated portfolio remains the same.

4.3.2 Active risk comparison

We now test the hypothesis of equal tracking errors for the 26 factor combinations and three portfolio construction methodologies presented in the previous section. We show the differences in the logarithmic tracking error and the multiple hypothesis adjusted p-values in Figure 5.

[Figure 5 about here.]

In contrast to Figure 4, there are not only positive but also negative (significant) differences

[13] We thank the referee for pointing us into this direction.

between integrated and mixed approach. Hence, on average, both approaches now have similar active risk and we are able to conduct a fair reward to risk analysis comparison in a next step.

4.3.3 Hypothesis testing

Adjusting for multiple hypothesis, we test the null hypothesis that the Sharpe ratio, variance, or information ratios are equal for the 26 style combinations and the three portfolio construction methodologies TER, BW, and TE from the set $\mathcal{P}$. Results are summarized in Figure 6.

[Figure 6 about here.]

For the Sharpe ratio in the top chart, we find no significant difference for any of the 78 combinations tested. We also find that many of the differences decrease and turn to negative numbers. For example, the BW approach shows only in three out of 26 combinations an improvement in the Sharpe ratio, while we found five positive significant Sharpe ratios in our first try in Section 4.1.

For the variance in the middle chart we find that 22 in the TER, four in the BW and eleven style combinations in the TE portfolio construction method show a significantly lower variance over time. In only seven of the 78 tested combinations, we observe a higher variance over time.

For the information ratio in the bottom chart we find for the TER methodology four out of 26 combinations with a higher information ratio in the integrated approach. In the BW methodology we find the combination *VL* to offer a minor improvement in the information ratio, while the other 25 combinations show lower information ratios. On the other hand, for the TE approach we see 19 of the 26 combinations to offer a higher information ratio. We can reject the null hypothesis that the two approaches are equal only for the five factor combination *VWCRL* of the BW portfolios. This combination shows a significant lower information ratio from June 1963 to December 2016.

We conclude that the significant improvements in the Sharpe ratio from Section 4 were due to the different level in active risk of the integrated to the mixed approach in long-only style investing. When we adjust the mixed approach such that it exhibits the same level in active risk, we can no longer reject the null hypothesis that the Sharpe ratio or information ratio is higher in the integrated approach. We even find one negative significant difference in the information ratio.

4.4 Asset pricing tests

To gain some further intuition about the differences of the mixed and integrated approach, we ask whether the return differences can be explained by the risk factors themselves. We therefore run for every combination in (10) and for the three portfolio construction methodologies TER, BW, and TE in $\mathcal{P}$ the following regression:

$$r_{int,t} - r_{mix,t} = \alpha + \beta_M M_t + \beta_S S_t + \beta_V V_t + \beta_R R_t + \beta_C C_t + \beta_W W_t + \beta_L L_t + \epsilon_t, \tag{11}$$

where $r_{int,t} - r_{mix,t}$ corresponds to the difference in monthly returns between the integrated and mixed approach, M is the market return, S is the small minus big factor of Fama and French (1992), and V, R, C, W, and L are the return differences of the upper tercile compared to the lower tercile within our equity universe defined in Section 3.

[Figure 7 about here.]

The parameter estimations and t-values are presented in Figure 7.[14] For the alpha coefficient, we cannot observe a consistent pattern in the t-statistics. To provide an explanation, we recall our stylized example in Section 2.1.1. On the one hand, the integrated approach increases the exposure to factor returns by penalizing negative characteristics. This property follows from aggregating the characteristics on security level, which keeps us from buying stocks with highly negative characteristics in one of the factors. For example, stocks B and D are not included in the integrated portfolio, because they include a negative factor score in one of the styles. On the other hand, the integrated approach decreases the sensitivity to factor returns by avoiding stocks that have diverging style exposures. For instance, stock B, which is highly sensitive to the momentum value factor, and stock D, which is highly sensitive to the value factor, are not included in the integrated approach. But stock C, with less pronounced style characteristics, is included. The insignificant alpha coefficients provide evidence that these two effects neutralize each other.

[14] We use HC3 t-values. The HC3 is a version of the significance tests based on a heteroscedasticity consistent covariance matrix (HCCM), which are consistent even in the presence of heteroscedasticity of an unknown form. We highlight t-values above 1.96 in green and below -1.96 in red. Moreover, we cap t-values above and below five to achieve a better overview.

When analyzing the different factor sensitivities, we observe that the sensitivity to the market factor M is low and mostly negative for combinations with three and more factors. The S factor, which can be seen as a proxy for illiquidity, is mostly negative with high negative t-values. This implies that the integrated approach loads less on liquidity risk, which may serve as an explanation of the lower expected returns in the long run of the integrated approach. For the factors V, R, C, and W, we find no clear pattern for the sensitivities. In contrast, for L we find a consistent positive sensitivity on the part of combinations with more than three factors, and large t-values. This is in line with the findings in Jivraj et al. (2016), who find high sensitivities to the low volatility factor when comparing the integrated and mixed approach for the style combination of value, momentum, low volatility, and quality. The high sensitivity to the low volatility factor L and the low realized risks of the low volatility factor over the analyzed time period are an explanation for the generally lower risk numbers of the integrated approach. The R-squared of the regressions increases with the number of factors considered, and obtain very high levels with an average R-squared of 0.30 for all style combination and portfolio construction techniques.

5 Turnover analysis

There are some limitations to our analysis. The first concerns the portfolio construction process. There are many ways of constructing a factor portfolio. We have focused on the most natural choices, the tercile portfolios being weighted by market capitalization as well as the benchmark-orientated approaches of Bender and Wang (2016) and Fitzgibbons et al. (2016). These constructions are close to the equilibrium portfolio of the CAPM and thereby minimize illiquidity issues. The second concern relates to potential selection biases. We tried to reduce such a bias by analyzing different factors and combinations as well. Yet, the set of factors and their definition was not know at the beginning of our analysis in 1963. Since the selection bias increases with the number of factors considered, results that depend on a large number of factors must be taken with caution.

A third concern, which is highly relevant from a practical viewpoint, is the turnover of the strategies. Figure 8 illustrates the turnover of the three portfolio construction methods with similar active risk for both the integrated and mixed approach with netting (mix - net) and without netting (mix

- gross). The mixed approach with netting views the turnover as if the single style portfolios are managed in one single mixed portfolio. For the mixed approach without netting, we calculate the turnover as if the single style portfolios are managed individually.

[Figure 8 about here.]

We observe that the integrated approach offers in general a lower (green) turnover compared to the mixed approach. When we compare the turnover reduction in the mixed approach with and without netting, we see that the effect is higher in the benchmark-orientated approaches and smaller in the decile approach. Finally for the style combinations including 'momentum' (W), the turnover in any of the portfolio constructions is substantially increased, while the combination of the quality type style factors 'robust' (R), 'investment' (C), and 'low volatility' (L) show much lower turnover over time.

We now test whether turnover, and therefore trading costs, have an impact on our previous results. We estimate transaction costs by starting with a one percent one-way transaction cost from 1963 to 1975. After the May Day in 1975 and the deregulation of the commission fees,[15] we decrease the transaction costs from 1976 to 2016 at an exponential decay with a mean lifetime of twelve years. This results in similar cost levels as used in different studies, such as, e.g., Keim and Madhavan (1998) and Jones (2002). Moreover, in 2016 the resulting transaction costs are 0.033 percent, which corresponds to the bid-ask spreads of US index funds at that time.[16]

[Figure 9 about here.]

Figure 9 shows the difference in the monthly Sharpe (top) and information ratio (bottom) together with the adjusted p-values of the integrated and mixed approach with netting. Due to the higher turnover of the mixed approach, we observe significant Sharpe ratio (information ratio) differences in the combinations WL, VWL, and $VWCL$ (WL and VWL) in the TE portfolios. Also, the reward to risk figures in the TER and BW increase. However, they are still negative or show high adjusted

[15] On May 1, 1975, brokerages were allowed to charge varying commission rates. Prior to this change, all brokerages charged the same price for stock trades.

[16] The bid-ask spread of index funds corresponds to the expected transaction costs in order to protect current investors from new subscriptions or redemptions.

p-values. Due to deregulation and higher volumes, commission fees and slippage decrease steadily over time. Therefore, we are also interested in the impact of trading costs for the more recent period.

[Figure 10 about here.]

Figure 10 shows the same tests for the period June 1993 to December 2016, with trading costs starting at 0.23 percent.[17] Strikingly, we find that for the period after June 1993, there is no evidence to reject the null hypothesis that the two approaches have the same return to reward ratio. The same conclusion holds for the mixed approach without netting. Hence, our analysis of transaction costs shows, due to the lower turnover, that the integrated approach may well be the better choice, if transaction costs are high. However, with the substantial decrease of these costs over the last decades, this advantage has eroded.

6 Conclusion

We rigorously study the difference in returns between the mixed and integrated approaches to long-only style investing. In the US stock market from 1963 to 2016, we analyze the 26 possible combinations of five styles: value, robustness, investment, momentum, and low volatility under the three portfolio construction methodologies of Fama and French (1992), Bender and Wang (2016) and a target tracking error optimization suggested in Fitzgibbons et al. (2016). While the previous literature concentrates on simple performance comparisons for arbitrary factor combinations and portfolio constructions, we apply a robust statistical testing framework. In contradiction to recent findings and the general tenor in the finance industry, we cannot support the hypothesis that the integrated approach leads to superior reward to risk ratios for any of the 26 tested factor combinations and portfolio construction methods.

We further find evidence that the integrated approach shows lower variances over time. In contrast to previous literature that mostly concentrates on a shorter and more recent time horizon, we find that the lower risk is, on average, associated with lower returns. For the integrated approach, we find a high sensitivity to the low volatility anomaly. By aggregating style information at security level,

[17] This breakpoint is also of interest due to the publication of the Fama-French three-factor model at this time and the studies of Bender and Wang (2016) and Fitzgibbons et al. (2016), which analyze data from 1993 onwards.

the integrated approach reduces risks and avoids extreme stocks that exhibit a high sensitivity to only a few (or only one) styles.

Our results confirm that, when naively tested, some factor combinations show superior return to risk ratios over specific periods. But when we apply a multiple hypothesis framework, we must conclude that none of the differences are significant. This conclusion also holds when we adjust for transaction costs. Given the increasing computational power for conducting multiple backtests and given the fact that financial institutions have incentives to deliver extraordinary results, it is crucial to apply the most advanced statistical testing frameworks. Ignoring the available tools can lead to hasty conclusions and mis-allocation of capital to investment strategies that are false discoveries.

References

Aharoni, Gil, Bruce Grundy, and Qi Zeng, 2013, Stock returns and the miller modigliani valuation formula: Revisiting the fama french analysis, *Journal of Financial Economics* 110, 347–357.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259–299.

Asness, Clifford S., and Andrea Frazzini, 2013, The devil in HML's details, *Journal of Portfolio Management* 39, 49–68.

Asness, Clifford S., Andrea Frazzini, and Lasse H. Pedersen, 2014, Quality minus junk, *Available at SSRN 2312432* .

Bailey, David H., Jonathan M. Borwein, Marcos L. de Prado, and Qiji J. Zhu, 2014, Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance, *Notices of the AMS* 61, 458–471.

Baker, Malcolm, Brendan Bradley, and Jeffrey Wurgler, 2011, Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly, *Financial Analysts Journal* 67, 40–54.

Bender, Jennifer, and Taie Wang, 2016, Can the whole be more than the sum of the parts? Bottom-up versus top-down multifactor portfolio construction, *Journal of Portfolio Management* 42, 39–50.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of statistics* 1165–1188.

Blitz, David, 2015, Factor investing revisited, *Journal of Index Investing* 6, 7–17.

Blitz, David, and Pim Van Vliet, 2007, The volatility effect: Lower risk without lower return, *Journal of Portfolio Management* 34, 102–113.

Bonferroni, Carlo E., 1936, *Teoria statistica delle classi e calcolo delle probabilita* (Libreria internazionale Seeber).

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Clarke, Roger G., Harindra De Silva, and Steven Thorley, 2016, Fundamentals of efficient factor investing, *Financial Analysts Journal* 72, 9–26.

Conneely, Karen N., and Michael Boehnke, 2007, So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests, *American Journal of Human Genetics* 81, 1158–1168.

Daniel, Kent D., and Tobias J. Moskowitz, 2015, Momentum crashes, *Journal of Financial Economics* forthcoming.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2009, Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?, *Review of Financial Studies* 22, 1915–1953.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fitzgibbons, Shaun, Jacques Friedman, Lukasz Pomorski, and Laura Serban, 2016, Long-only style investing: Don't just mix, integrate, Integrate (June 29, 2016), AQR Capital Management, LLC.

Fraser-Jenkins, Inigo, Alix Guerrini, Alla Harmsworth, Mark Diver, Sarah McCarthy, Robertas Stancikas, and Maureen Hughes, 2016, Global quantitative strategy: How to combine factors? It depends why you are doing it, Global Quantitative Strategy (September 14, 2016), Sanford Bernstein.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, ...and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Haugen, Robert A., and Nardin L. Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of Financial Economics* 41, 401–439.

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 65–70.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Jensen, Michael C., Fischer Black, and Myron S. Scholes, 1972, *The capital asset pricing model: Some empirical tests* (Praeger Publishers Inc).

Jivraj, Farouk, David Haefliger, Zein Khan, and Benedict Redmond, 2016, Equity multi-factor approaches: Sum of factors vs. multi-factor ranking, QIS Insights (September 16, 2016), Barclays Investment Bank.

Jones, Charles M., 2002, A century of stock market liquidity and trading costs, *Graduate School of Business, Columbia University* .

Keim, Donald B, and Ananth Madhavan, 1998, The cost of institutional equity trades, *Financial Analysts Journal* 50–69.

Ledoit, Olivier, and Michael Wolf, 2004, Honey, i shrunk the sample covariance matrix, *The Journal of Portfolio Management* 30, 110–119.

Ledoit, Olivier, and Michael Wolf, 2008, Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Ledoit, Olivier, and Michael Wolf, 2011, Robust performances hypothesis testing with the variance, *Wilmott* 55, 86–89.

Leippold, Markus, and Harald Lohre, 2012a, Data snooping and the global accrual anomaly, *Applied Financial Economics* 22, 509–535.

Leippold, Markus, and Harald Lohre, 2012b, International price and earnings momentum, *The European Journal of Finance* 18, 535–573.

Leippold, Markus, and Harald Lohre, 2014, The dispersion effect in international stock returns, *Journal of Empirical Finance* 29, 331 – 342.

Leippold, Markus, and Roger Rueegg, 2017, The mixed vs the integrated approach to style investing: Much ado about nothing?, *European Financial Management* forthcoming.

Novy-Marx, Robert, 2013, The other side of value: The gross profitability premium, *Journal of Financial Economics* 108, 1–28.

O'Brien, Robert M., 2007, A caution regarding rules of thumb for variance inflation factors, *Quality & Quantity* 41, 673–690.

Patton, Andrew, Dimitris N. Politis, and Halbert White, 2009, Correction to "automatic block-length selection for the dependent bootstrap" by D. Politis and H. White, *Econometric Reviews* 28, 372–375.

Politis, Dimitris N., and Halbert White, 2004, Automatic block-length selection for the dependent bootstrap, *Econometric Reviews* 23, 53–70.

Romano, Joseph P., and Michael Wolf, 2005a, Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* 100, 94–108.

Romano, Joseph P., and Michael Wolf, 2005b, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P., and Michael Wolf, 2016, Efficient computation of adjusted p-values for resampling-based stepdown multiple testing, *Statistics & Probability Letters* 113, 38–40.

Figure 1. Integrated vs. mixed approach: Sharpe ratios

This figure presents the Sharpe ratios of the mixed and integrated approach in gray and difference between integrated and mixed approach in green (positive) or red (negative). The analyzed factors are: 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). E.g., the combination of 'value', 'momentum' and 'low volatility' is indicated by VWL. The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). Portfolios are rebalanced monthly from June 1963 to December 2016.

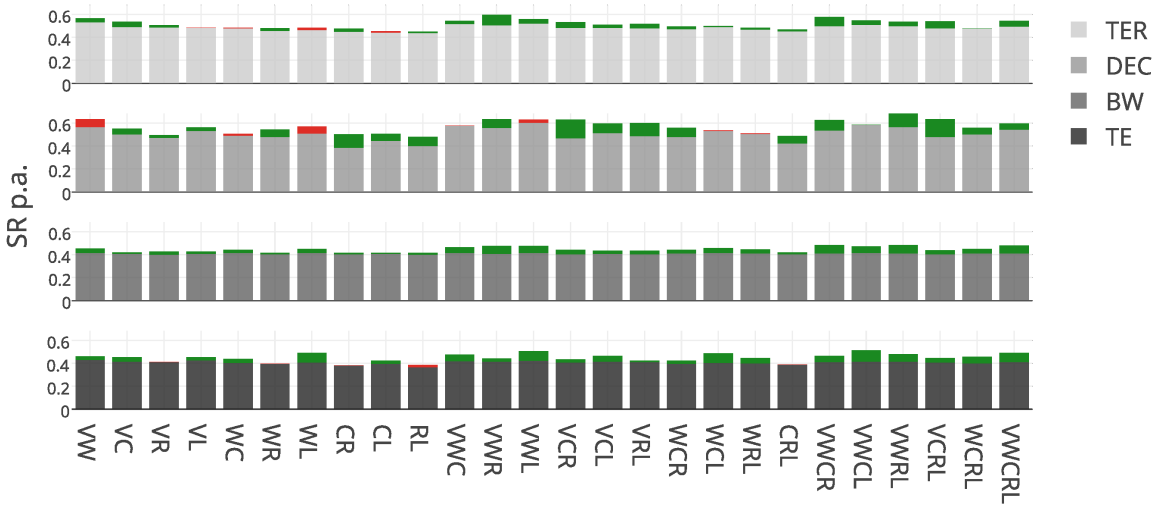


Figure 2. Robust Sharpe ratio: single hypothesis test

This figure presents the comparison of the Sharpe ratios of the integrated and mixed approach to long-only style investing. The analyzed factors are value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe ratio (SR diff) in bars and the p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) for a block size of five in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate. The data starts in June 1963 and ends in December 2016.

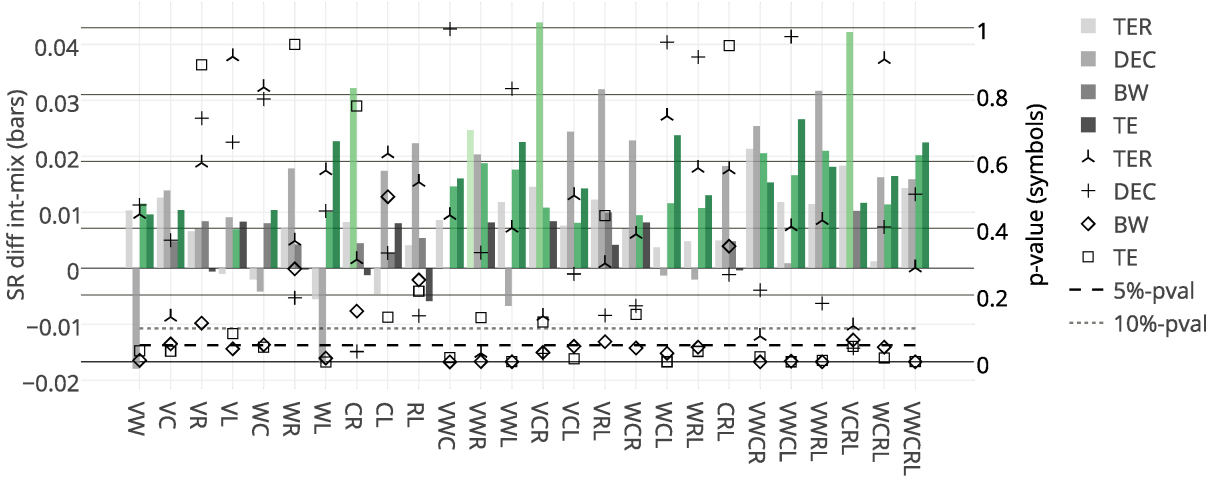


Figure 3. Robust Sharpe ratio: multiple hypothesis test

This figure presents the comparison of the Sharpe ratios of the integrated and mixed approach to long-only style investing. The analyzed factors are value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe ratio (SR diff) in bars and the p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) for a block size of five adjusted by the multiple hypothesis framework of Romano and Wolf (2016) in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate. The data starts in June 1963 and ends in December 2016.

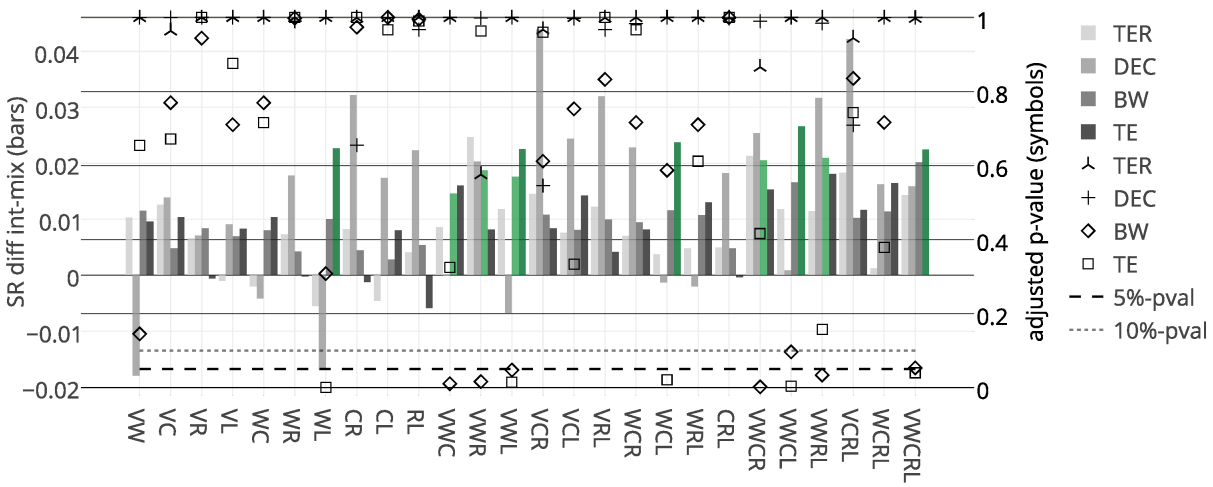


Figure 4. Multiple hypothesis test: variance, information ratio, tracking error

This figure presents the differences in the logarithmic variance in bars (VR diff) and the adjusted p-values of the robust variance test of Ledoit and Wolf (2011) in symbols in the top chart; differences in the information ratio relative to the market-cap-weighted benchmark (IR diff) in bars and the adjusted p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) in symbols in the middle chart; differences in the logarithmic tracking error in bars (TE diff) and the ajdusted p-values of the robust tracking error test of Ledoit and Wolf (2011) in symbols in the bottom chart. The single hypothesis p-values are adjusted for the number of tries by the multiple hypothesis framework of Romano and Wolf (2016). The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). The portfolios are rebalanced monthly from June 1963 to December 2016. The analysis is based on the monthly excess returns above the one-month Treasury bill rate.

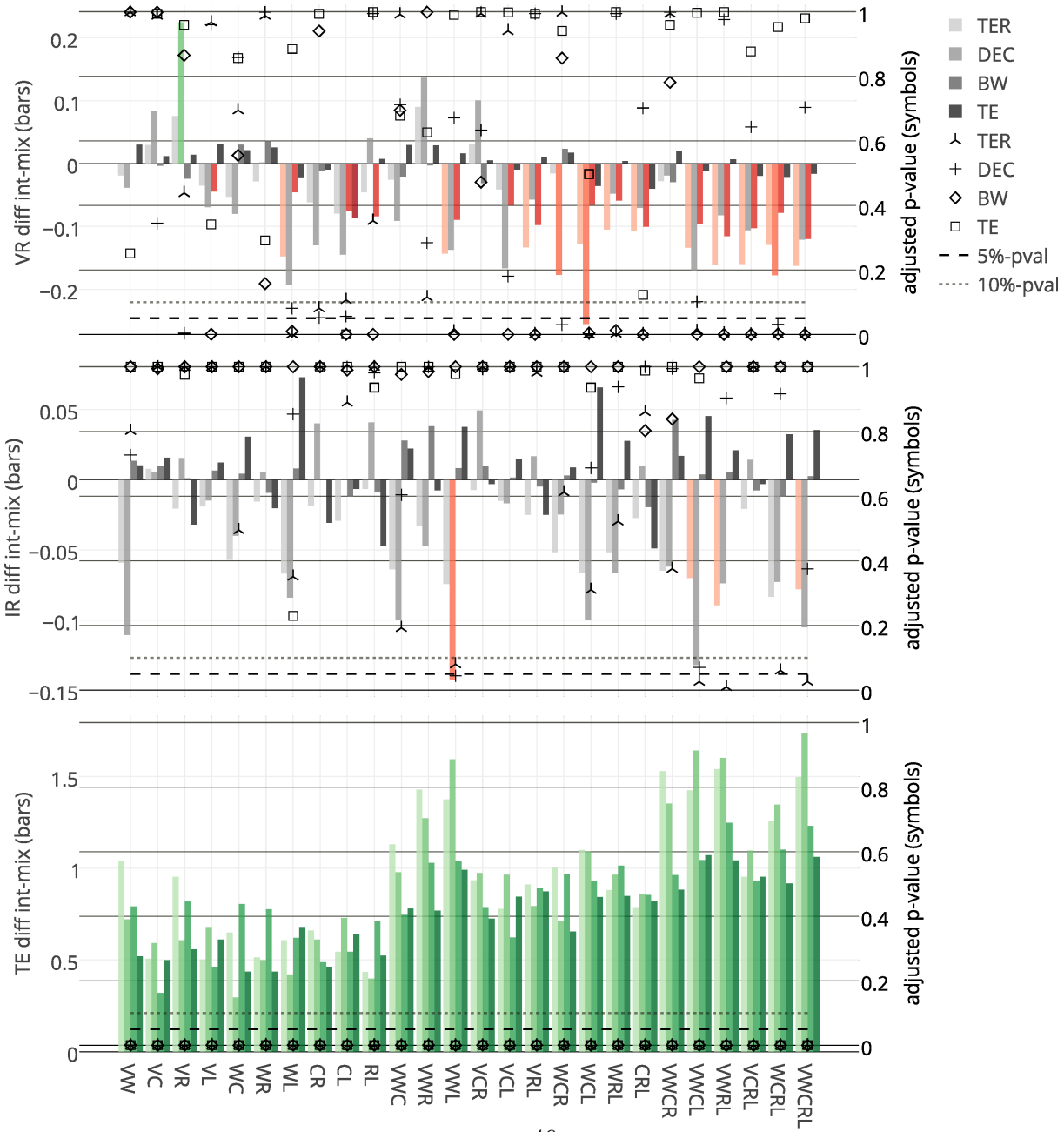


Figure 5. Same active risk: robust tracking error test

This figure presents the comparison of the tracking error (TE diff) of the integrated approach with those of the mixed approach to long-only style investing. The analyzed factors are value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The portfolio construction methodologies tested are the TER, BW, and TE. The mixed portfolios is constructed with higher concentrated style portfolios to achieve a similar active risk compared to the integrated approach.The portfolios are rebalanced monthly. We show the difference in the logarithmic tracking error (TE diff) in bars and the p-values of the robust variance test of Ledoit and Wolf (2008) adjusted by the multiple hypothesis framework of Romano and Wolf (2016) for a block size of five in symbols. The analysis is based on the monthly excess returns above the market-cap-weighted benchmark. The analysis starts in June 1963 and ends in December 2016.

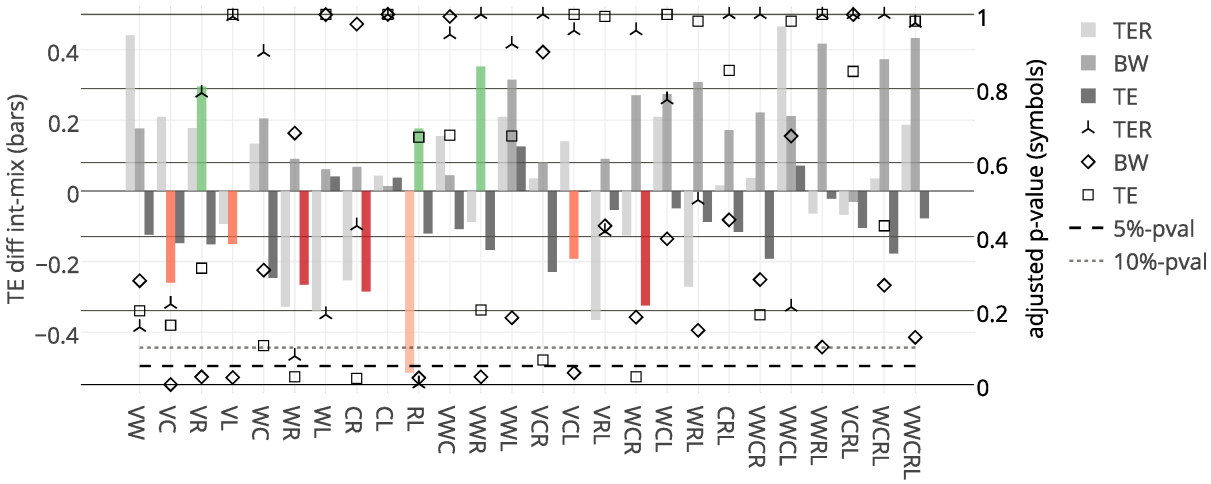


Figure 6. Same active risk: Sharpe ratio, variance, information ratio

This figure presents the difference in the Sharpe ratio (SR diff) and the adjusted p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) in the top chart; differences in the logarithmic variance in bars (VR diff) and adjusted p-values of the robust variance test of Ledoit and Wolf (2011) in the middle chart; difference in the information ratio relative to the market-cap-weighted benchmark (IR diff) in bars and the adjusted p-values of the robust Sharpe ratio test in the bottom chart. The single hypothesis p-values are adjusted for the number of tries by the multiple hypothesis framework of Romano and Wolf (2016). The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are TER, BW, and TE. The analysis is based on the monthly excess returns above the market-cap-weighted benchmark. The analysis starts in June 1963 and ends in December 2016.

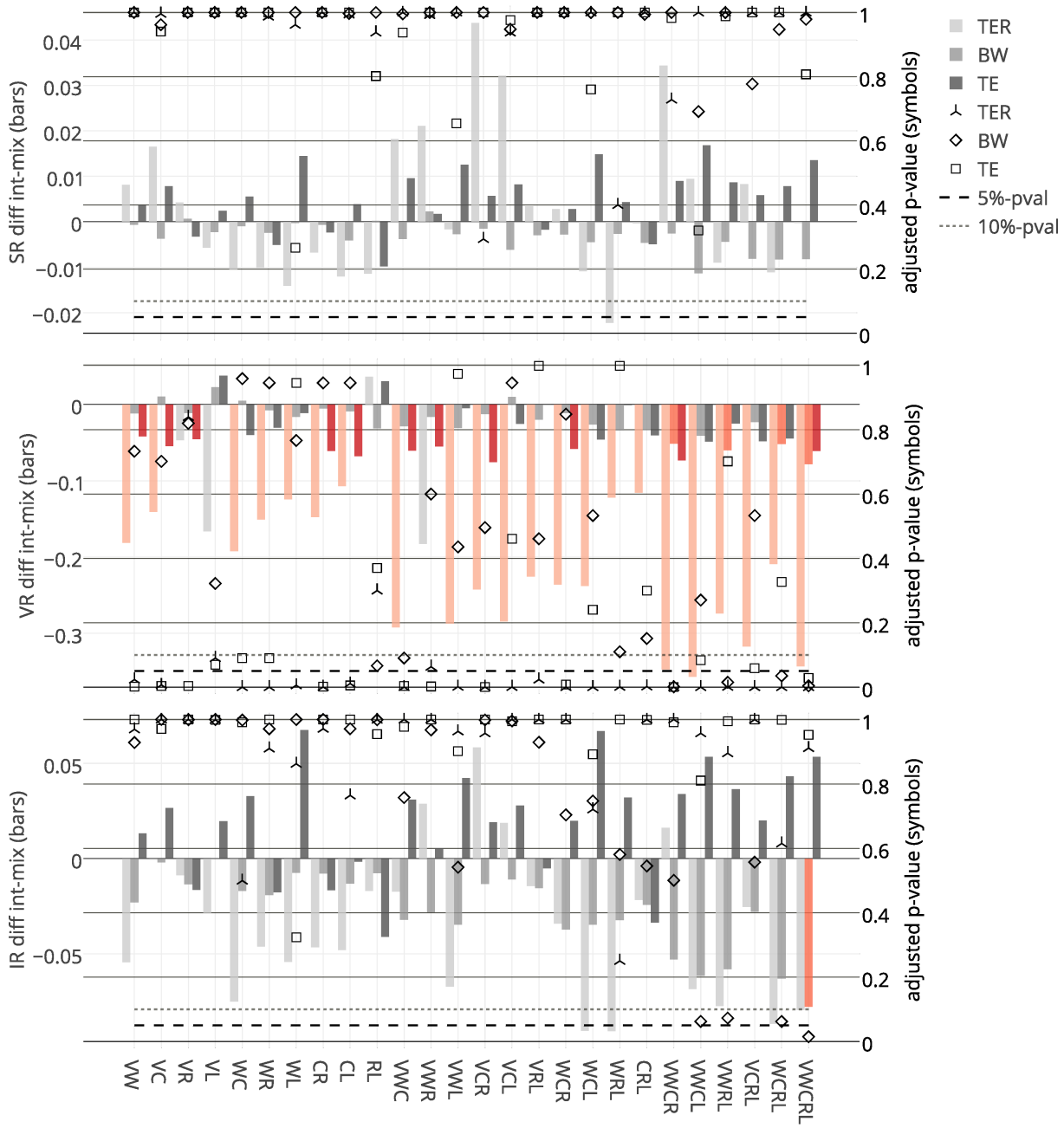


Figure 7. Asset pricing tests

This figure presents the ordinary least squares regression results with data from June 1963 to December 2016. The dependent variable is the difference in the monthly returns of the integrated approach from those of the mixed approach in long-only style investing. The independent variables are the market portfolio (MKT), the small minus big (SMB), the value (V), the profitability (R), the conservative (C), the momentum (W), and the low volatility (L) factors. Factor returns are calculated by the difference in the performance of the highest to the lowest tercile. Except for the factor small minus big, which is defined as in Fama and French (1992), we only use the big universe to calculate the factor returns. We report the HC3 t-values (t) for the 26 possible factor combinations of V,R,C, W, and L as well as for the tercile (TER), Bender and Wang (2016) (BW), and target tracking error optimization (TE) portfolio construction. HC3 test statistics above (green) and below (red) 1.96 are highlighted and the test statistics are truncated at +/- five for a better overview.

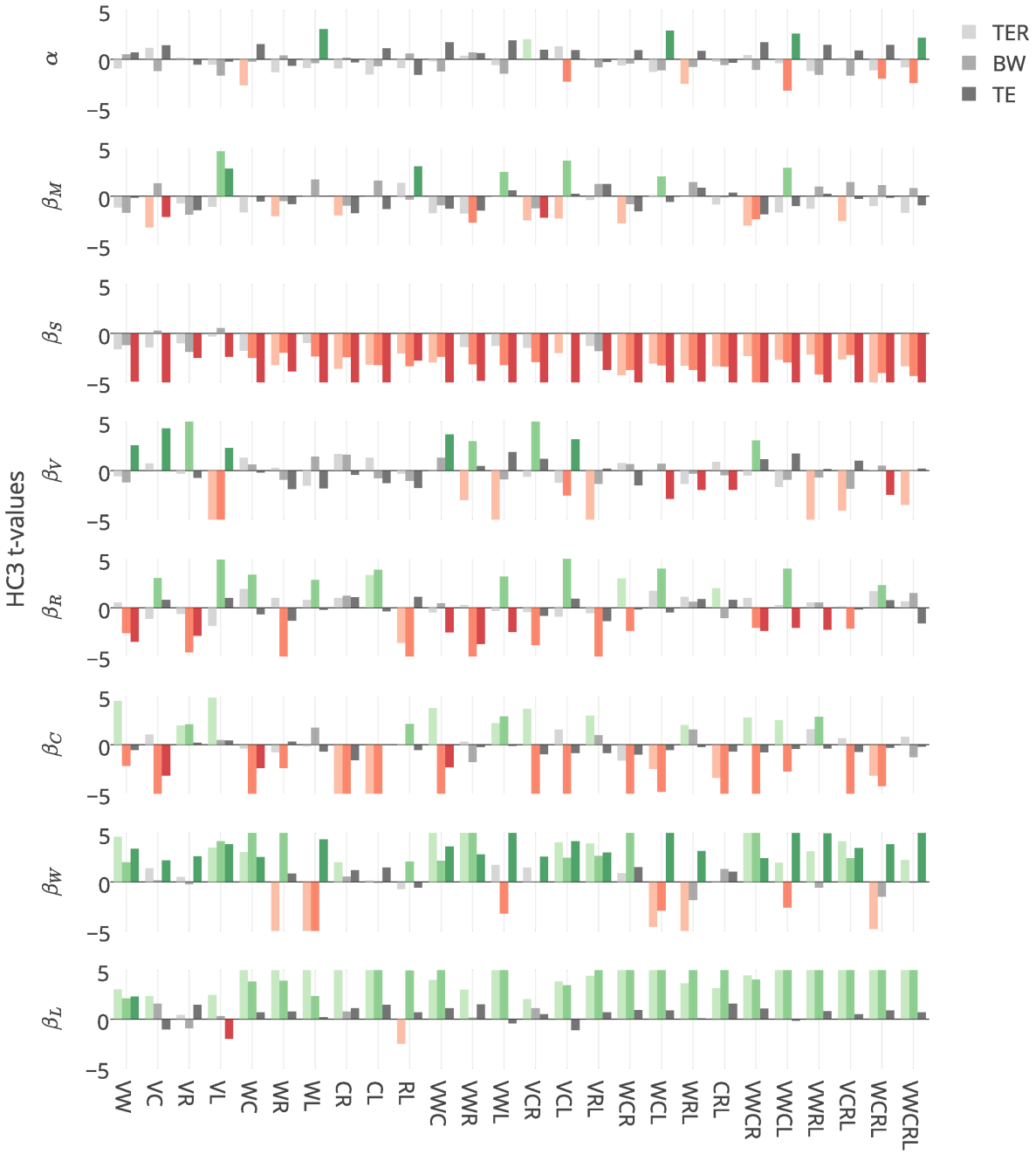


Figure 8. Turnover Analysis

This figure presents the monthly average turnover of the integrated (integrate), mixed approach with netting (mix - net) and mixed approach without netting (mix - gross) from June 1963 to December 2016. The portfolio construction methodologies tested are the TER, BW, and TE. The portfolios are constructed in a way that the active risk of both, the integrated and mixed approach, are similar over time. A lower (higher) turnover of the integrated approach compared to the mixed approach with netting is highlighted in green (red).

Figure 9. Multiple hypothesis test with transaction costs: 1963 - 2016

This figure presents the comparison of the Sharpe (top) and information ratios (bottom) of the integrated approach with those of the mixed approach to long-only style investing. The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are TER, BW, and TE. The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe (SR diff) and information ratio (IR diff) in bars and the multiple hypothesis p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) adjusted by the multiple hypothesis framework of Romano and Wolf (2016) for a block size of five in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate including trading costs. The analysis starts in June 1963 and ends in December 2016.

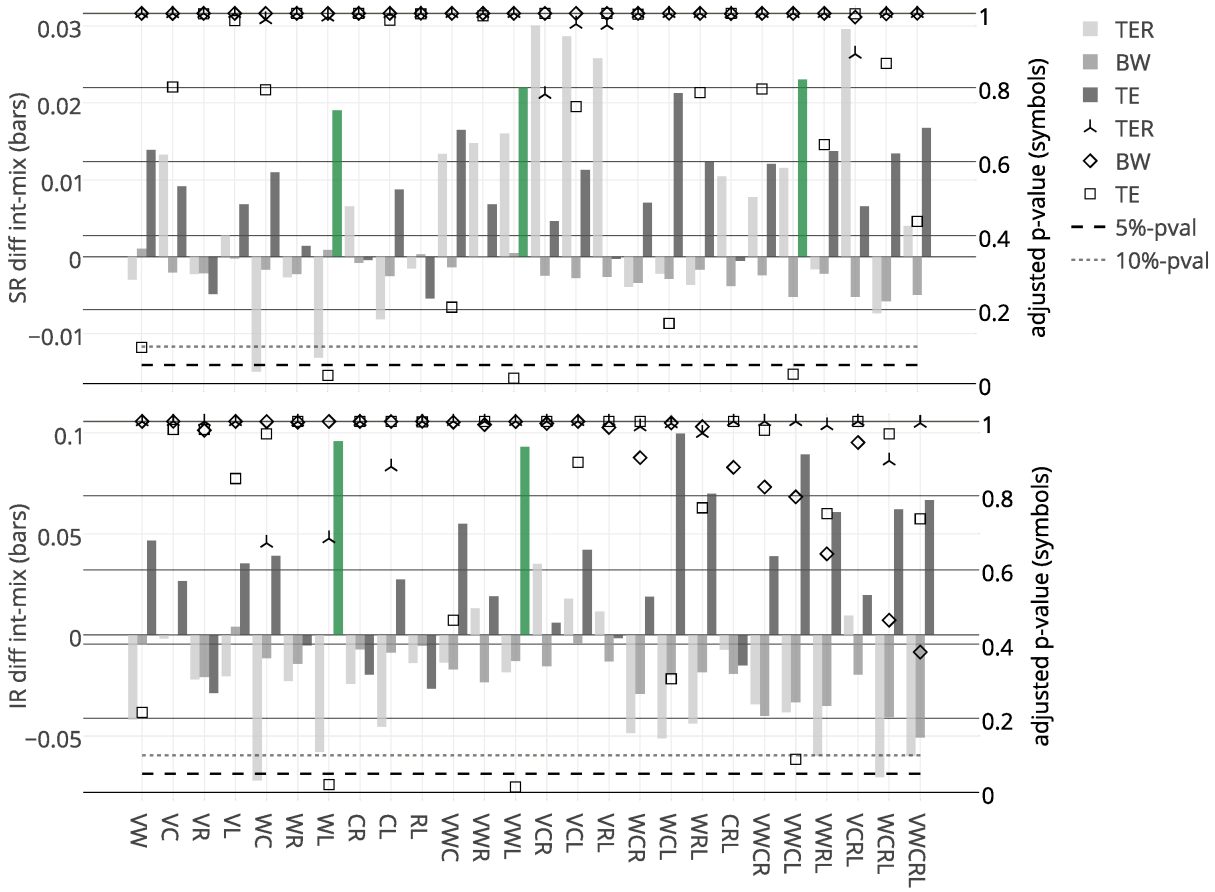


Figure 10. Multiple hypothesis test with transaction costs: 1993 - 2016

This figure presents the comparison of the Sharpe (top) and information ratios (bottom) of the integrated approach with those of the mixed approach to long-only style investing. The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are the tercile (TER), Bender and Wang (2016) (BW), and target tracking error optimization (TE). The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe (SR diff) and information ratio (IR diff) in bars and the multiple hypothesis p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) adjusted by the multiple hypothesis framework of Romano and Wolf (2016) for a block size of five in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate including trading costs. The analysis starts in June 1993 and ends in December 2016.

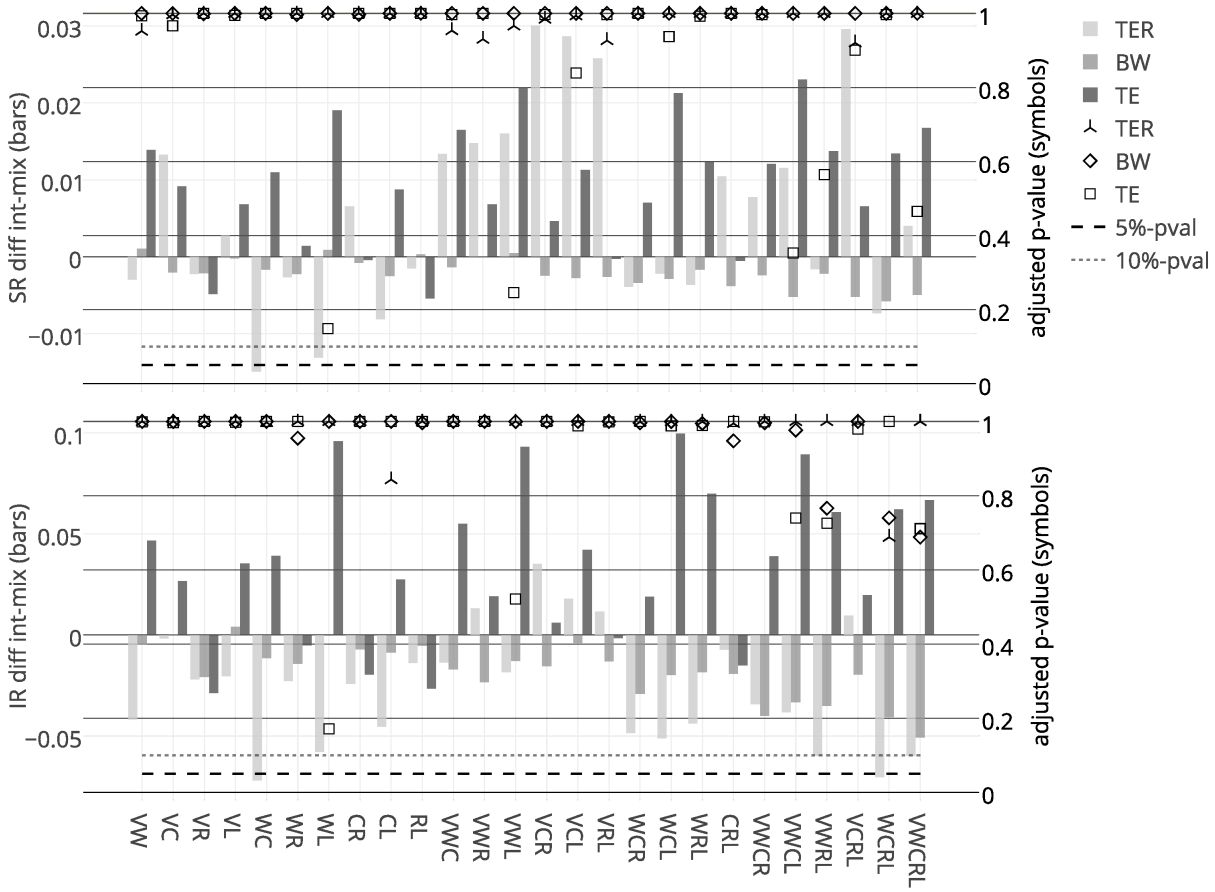


Table 1. Multi-factor ETFs

This table presents the most well known multi-factor ETFs as of end of August 16. Data compiled by Bloomberg and ETF.com and ordered by assets under management (AuM) as of end of August 16.

Name	Asset Manager	AuM	Inception	Approach
Goldman Sachs ActiveBeta U.S. Large Cap Equity ETF	Goldman Sachs	US $1.15B	01/28/15	mix
FlexShares Morningstar US Market Factor Tilt Index Fund	FlexShares	US $842.43M	09/16/11	integrate
John Hancock Multifactor Large Cap ETF	John Hancock	US $236.27	09/28/15	integrate
State Street Multi-Factor Global Equity Fund	State Street	US $126.39	09/30/14	mix
iShares Edge MSCI Multifactor USA ETF	iShares	US $110.03M	04/30/15	integrate
JPMorgan Diversified Return U.S. Equity ETF	JP Morgan	US $81.15M	09/29/15	integrate
The Global X Scientific Beta US ETF	Global X	US $67.18M	05/12/15	mix
Franklin LibertyQ Global Equity ETF	Franklin	US $26.19M	06/01/16	integrate
ETFS Diversified-Factor U.S. Large Cap Index Fund	ETF Securities	US $7.82M	01/28/15	mix

Table 2. Stylized example

This table presents a stylized example of the calculations of the weights of the mixed and integrated portfolios. We report the market capitalization (mc), the factor values of value (V) and momentum (W), the weights in the factor portfolio of value (w_V) and momentum (w_W), the score for value (s_V), the score for momentum (s_W), the aggregated scores for the combination (s_{VW}), and the resulting weights for the mixed portfolio (w_{mix}) and the integrated portfolio (w_{int}).

	mc	ϕ_V	ϕ_W	s_V^{rank}	s_W^{rank}	$s_{\text{agg}}^{\text{rank}}$	w_V	w_W	w_{mix}	w_{int}
Stock A	10.50	**0.51**	**0.09**	**0.77**	**0.77**	**0.77**	**0.23**	**0.37**	**0.30**	**0.35**
Stock B	5.75	0.22	**0.14**	0.11	**0.88**	0.50	0.00	**0.20**	**0.10**	0.00
Stock C	7.12	0.48	0.02	0.66	0.55	**0.61**	0.00	0.00	0.00	**0.24**
Stock D	22.87	**0.82**	-0.09	**0.88**	0.11	0.50	**0.50**	0.00	**0.25**	0.00
Stock E	7.72	0.32	-0.21	0.44	0.00	0.22	0.00	0.00	0.00	0.00
Stock F	1.15	0.14	0.07	0.00	0.66	0.33	0.00	0.00	0.00	0.00
Stock G	15.72	0.31	0.01	0.33	0.44	0.38	0.00	0.00	0.00	0.00
Stock H	50.91	0.28	-0.07	0.22	0.22	0.22	0.00	0.00	0.00	0.00
Stock I	12.51	**0.97**	**0.22**	**1.00**	**1.00**	**1.00**	**0.27**	**0.43**	**0.35**	**0.42**
Stock J	25.26	0.41	-0.02	0.55	0.33	0.44	0.00	0.00	0.00	0.00

Table 3. Multiple hypothesis testing: Momentum strategy

This table presents the comparison of the Sharpe ratios for 20 momentum strategies of IBM compared to the buy and hold strategy. The momentum strategies invest in IBM for the next month if the xth most recent month was positive. Otherwise, we step out of the stock for the next month. We show the strategy for $x = 1, ..., 20$. The out-of-sample backtest starts in June 1963 and ends in December 2014. We show the annualized return (Ret p.a.), the annualized volatility (Vol p.a.), the monthly Sharpe Ratio difference (SR-diff), the bootstrapped p-value of Ledoit and Wolf (2008) with a block size of two, the p-values adjusted by the frameworks of Bonferroni (Bonf), Holm (1979) (Holm), Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) (BHY), and Romano and Wolf (2016) (RW). The analysis is based on monthly excess returns above the one-month Treasury bill rate. *** denotes significance at the 0.01 level; ** denotes significance at the 0.05 level; and * denotes significance at the 0.1 level.

x =	1	2	3	4	5	6	7	8	9	10
Ret p.a.	0.051	0.023	0.021	0.008	0.010	-0.008	0.023	0.012	0.041	0.029
Vol p.a.	0.163	0.164	0.158	0.159	0.160	0.176	0.173	0.160	0.166	0.175
SR-diff	0.030	-0.018	-0.021	-0.044	-0.041	-0.069	-0.019	-0.037	0.012	-0.010
pval	0.347	0.576	0.507	0.170	0.203	0.015**	0.511	0.242	0.708	0.742
Bonf	1.000	1.000	1.000	1.000	1.000	0.292	1.000	1.000	1.000	1.000
Holm	1.000	1.000	1.000	1.000	1.000	0.277	1.000	1.000	1.000	1.000
BHY	0.645	0.752	0.731	0.485	0.506	0.146	0.731	0.537	0.752	0.752
RW	0.985	0.994	0.994	0.904	0.933	0.230	0.994	0.949	0.996	0.996

x =	11	12	13	14	15	16	17	18	19	20
Ret p.a.	0.006	0.042	0.004	0.019	-0.013	0.071	0.040	0.026	0.017	0.005
Vol p.a.	0.171	0.171	0.163	0.159	0.175	0.165	0.176	0.163	0.168	0.164
SR-diff	-0.047	0.012	-0.051	-0.026	-0.077	0.062	0.009	-0.013	-0.028	-0.049
pval	0.127	0.681	0.100	0.449	0.008***	0.053*	0.752	0.691	0.355	0.138
Bonf	1.000	1.000	1.000	1.000	0.168	1.000	1.000	1.000	1.000	1.000
Holm	1.000	1.000	1.000	1.000	0.168	0.958	1.000	1.000	1.000	1.000
BHY	0.461	0.752	0.461	0.731	0.146	0.355	0.752	0.752	0.645	0.461
RW	0.848	0.996	0.818	0.993	0.141	0.580	0.996	0.996	0.985	0.857

Table 4. Factors' summary statistics

This table presents the annualized return (Ret p.a.), annualized volatility (Vol p.a.), Sharpe ratio (SR p.a.), and maximum draw-down (Max. Draw.) for the value-weighted monthly excess returns above the one-month Treasury bill rate. The period starts in June 1963 and ends in December 2016. The market includes all securities. Small: the securities below the NYSE market capitalization median; Big: the securities above the NYSE market capitalization. The other factors are the top (first-listed), respectively, bottom (second-listed), terciles in the big universe of securities of the following factors: 'value' (Value - Growth), 'robustness' (Robust - Weak), 'investment' (Conser. - Aggr), 'momentum' (Winner - Loser), and 'low volatility' (LowVol - HighVol).

1963 - 2016	Market	Small	Big	Value	Growth	Robust	Weak	Conser.	Aggr.	Winner	Loser	LowVol	HighVol
Ret p.a.	5.14	6.48	5.07	8.67	4.71	6.39	2.76	7.11	4.90	8.10	2.39	5.50	4.49
Vol p.a.	15.53	21.55	15.13	16.38	16.19	15.22	18.11	15.06	17.90	16.91	19.81	12.70	25.67
SR p.a.	0.33	0.30	0.34	0.53	0.29	0.42	0.15	0.47	0.27	0.48	0.12	0.43	0.17
Max. Draw.	55.23	73.83	55.44	54.11	58.60	53.04	78.59	47.54	63.72	50.39	72.71	48.90	78.84

Table 5. Variance inflation factors

This table presents the variance inflation factors of the following factors: value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The return series of each factor are computed as the value-weighted return of the upper tercile less the value-weighted return of the lower tercile of each factor for the big universe. We show the variance inflation factors for each of the 5 factors as independent variables. The dependent variable of the model is shown on the horizontal axis, while the independent variables are shown on the vertical axis.

	V	C	W	R	L
V	-	1.49	1.03	1.36	1.73
C	1.46	-	1.16	1.38	1.48
W	1.74	1.99	-	1.42	1.76
R	1.91	1.97	1.18	-	1.32
L	1.95	1.70	1.18	1.06	-

Is Active Investing a Zero-Sum Game?

Markus Leippold and Roger Rueegg

I have presented the paper at:

- Knowledge Transfer Seminar, October 2017, Zürich Cantonal Bank, Switzerland.
- UZH Brown Bag Doctoral Lunch Seminar, March 2018, Zürich, Switzerland.
- Frontiers of Factor Investing Conference, April 2018, Lancaster University, England.

It will be presented at:

- 16th Paris December Finance Meeting, December 2018, Paris, France.

Abstract

To study the hypothesis whether active investing is a zero-sum game, we analyze the alpha of active and index mutual funds from a global sample of more than 60,000 equity and fixed income funds. Using a new robust statistical test, we cannot reject this hypothesis for the vast majority of investment categories. We also find that the average active fund has less exposure to traditional risk factors, but higher sensitivity to alternative risk premia. Fund persistence and the impact of size and fees adds further support to the hypothesis.

1 Introduction

The emergence of index investing has led to a seemingly endless debate about the merits of active portfolio management. Many research papers, investors, and advisors place themselves in either the active or passive camp. The staunch defenders of active investing argue along the lines of Berk and Green (2004), who show that rational markets do not contradict the existence of skilled fund managers who consistently beat the market. They build their argument on a basic principle of economics: agents earn economic rents if, and only if, they have a competitive advantage. Hence, active investing is a zero-sum game after fees. Recently, Berk and van Binsbergen (2015) have provided empirical support for the claim that mutual fund managers do have skills.

In contrast, the proponents of passive investing argue along the lines of Fama and French (2010) in that the high fees of active management turn it into a negative-sum game after costs. Indeed, French (2008) and Fama and French (2010), among many others, provide ample evidence that actively managed US equity mutual funds underperform their multi-factor benchmark after fees. In their view, active investing is at most a zero-sum game before fees, but definitely not after fees. Consequently, over the recent years, we have witnessed a massive inflow of funds into index investing. These observations naturally drive us to question the value of active management.

According to the logic of Sharpe (1991)'s active management arithmetic, active investing is doomed in aggregate, as French (2008) puts it. However, to escape this seemingly irrefutable conclusion, Sharpe (1991) leaves a back door open by pointing out three potential flaws in his theory. First, passive managers might not be truly passive. Second, there might be substantial differences among active managers.[1] Third, the summary statistics of active managers might not truly represent the performance of the actively managed dollar. In our analysis, we shed light on these three potential pitfalls by first applying a factor analysis not only on active but also on index fund. Second, we account for the heterogeneity of active asset managers by differentiating between institutional and

[1]For example, Garleanu and Pedersen (2018) consider a model where managers with larger and more sophisticated investors are expected to outperform. Their theoretical model is supported, e.g., by the findings of Gerakos et al. (2016), who show that institutional investors outperform their strategy benchmarks after fees. Their analysis is based on self-reported but GIPS (Global Investment Performance Standard) compliant data, which still may inherit some biases, while our analysis is based on publicly available performance data. In addition, Pastor et al. (2015) argue that the higher competitions in big active mutual fund industries decrease the fund's ability to outperform passive benchmarks.

retail funds, equity and fixed income funds, geographical regions, and investment categories. Lastly, we analyze the value-weighted performance before and after fees, and we benchmark against multi-factor models and investable indexes. For this extensive study, we include 61,269 equity and fixed income funds that held USD 17.8 trillion assets under management by the end of 2016, thereby substantially increasing the power of our tests.[2]

Since our preliminary data analysis indicates that there are both serial and cross-sectional dependencies in our fund data, we have developed a robust test for the manager's alpha, defined as the excess return relative to an appropriate benchmark. Our test is robust in the sense that it takes into account potential serial dependence in mutual fund returns.[3] We can then use these robust test statistics for the alphas as input for the appropriate multiple hypothesis adjustments.

As Berk and van Binsbergen (2017) convincingly argue, there is no unique way to measure performance: it depends on the research question. If we want to assess the rationality of fund investors and the degree of competition in different markets, then the appropriate measure is the fund's net alpha. If, however, we want to measure the manager's skill, then the appropriate measure is the value-weighted gross alpha, or if we want to test whether active and index investing is a zero-sum game after costs, then we must use the value-weighted net alpha. Unlike most of the previous studies, which prominently focus on active US mutual funds without differentiating between retail and institutional funds, we use the richness of our dataset to explore the performance of active funds from many different perspectives.[4]

The choice of benchmark is just as critical as the measurement of performance. Typically, researchers use a well established multi-factor model to proxy for the alternative opportunity set available to investors. However, multi-factor models include long–short portfolios with often very high turnover, generating considerable transaction costs. Furthermore, also as argued by Berk and van

[2]Of these funds, 56,136 are actively managed, and 5,133 are index funds.

[3]In a simulation study, we find that the conventional inference techniques are liberal in rejecting the null hypothesis, while we observe still liberal but accurate empirical rejection probabilities for our robust alpha test based on block resampling. Other papers that conduct bootstrapped inference, such as Kosowski et al. (2006) and Fama and French (2010), sample one-period returns and, therefore, lose any information about the potential dependence over time. Hence, the simulation results convince us that our statistical framework is the appropriate choice to carry out our research task.

[4]Recently, Ferreira et al. (2013) analyze the performance of 16,316 open-end actively managed equity funds in 27 countries from 1997 to 2007. They find that equity mutual funds around the globe, in general, underperform. Banegas et al. (2013) focus on 4,200 European equity mutual funds and find that European equity funds outperform the market.

Binsbergen (2017), we might have a situation in which we are measuring the performance of a fund at a time when the fund manager would not have known about some factors, as they were identified only much later.

Nevertheless, Fama and French (2010) argue that benchmarking against multi-factor models leads to the same conclusions as benchmarking against index funds, because the value-weighted portfolio of index funds exhibits close to zero alphas. Although we agree with their arguments for the US equity market, our single fund analysis reveals that many index funds also exhibit negative alphas relative to a multi-factor benchmark, depending on the asset class and the market. Also, there is no general agreement on the factors that should be included in a benchmark, leading to a severe selection bias. Therefore, to measure fund performance, we abandon multi-factor models as a benchmark in favor of suitably defined investable benchmarks, which allow a fair comparison of active and index funds. Of course, we still need multi-factor models to understand the potentially different risk and style exposures of active and index funds.

Under the assumption that index investors try to replicate the market portfolio and believe in the efficient market hypothesis, we compare the average dollar weighted return of active investors that build an opinion with the average dollar weighted return of an investor who has no opinion about the securities within a certain investment category.[5] Using such an approach, we also take into account the costs of replicating the market portfolio that arise due to management fees, transaction costs, and we also guarantee that factors with a low market capitalization receive a lower weight.

In our preliminary analysis, we find significant evidence for serial and cross-sectional dependence in our mutual fund data. Therefore, we use a statistical framework with two key elements. First, we develop a robust statistical test for the mutual funds' alpha, which takes into account serial dependence. Second, we use these test statistics as input for the multiple hypothesis testing methods of Barras et al. (2010) and Romano and Wolf (2016), which are robust to the presence of (mild) cross-sectional dependence. In a simulation study, we find that the standard inference techniques are liberal in rejecting the null hypothesis, while we find accurate empirical rejection probabilities for

[5]Similarly, Fama and French (2010) use such an equilibrium accounting perspective to argue that the actively managed mutual fund industry does not cover the costs they impose on investors. However, they only concentrate on one portfolio, formed by value-weighting the funds in the "Active US Equity" category, and compare it to the market portfolio.

our block resampling based alpha test. Other papers that conduct bootstrapped inference, such as Kosowski et al. (2006) and Fama and French (2010), sample one-period returns and, therefore, lose any information on the potential serial dependence; nor do they allow for multiple hypotheses.[6]

When applying our robust multiple hypothesis alpha test to single funds against multi-factor benchmarks, we find that a large fraction of active equity mutual funds deliver negative alphas after fees. Therefore, we provide international evidence for the results of Fama and French (2010).[7] Surprisingly, however, when we apply the same tests to index funds, we find that they also show negative alphas after costs. While most of the literature concentrates on the equity market, we also conduct a multi-factor analysis of the fixed income mutual funds. In contrast to the equity market, we find that there is a substantial portion of active USD fixed income funds with a positive alpha.

Given that we observe substantially negative alphas for index funds in both the equity and fixed income markets, we question the plausibility of using multi-factor benchmarks.[8] Therefore, in a further step, we analyze the net alphas of single active funds against investable benchmarks. We find that most of the active funds exhibit zero alpha. While for institutional equity funds in the US, we find a negligible proportion of negative-alpha funds, this proportion is higher for retail equity funds. Again, we observe a large fraction of index funds with negative alphas under an investable benchmark. We suspect that this negative performance may be caused by the negative performance of small funds.

Since Berk and Green (2004) argue that fund size is a crucial element in the analysis of fund performance, we value-weight the alpha of active funds within the Morningstar investment categories. The multiple hypothesis test of Barras et al. (2010) cannot be applied to value-weighted alphas. Therefore, we switch to the method of Romano and Wolf (2016). We find that there are significant negative alphas after cost for the institutional "US Equity Large Cap Blend" and retail "Canada Fixed Income" categories. This finding corroborates the conclusion of Fama and French (2010) for these specific investment categories. However, for all of the other categories, our results support Berk and

[6]The Matlab code of the robust statistical framework for the alpha is available from the authors on request.

[7]In particular, we find the highest proportion of funds with positive alphas against the multi-factor benchmark for institutional investors to be in Europe and Japan, which confirms the findings of Banegas et al. (2013).

[8]Index funds should have zero alpha on average if the multi-factor benchmark were appropriate. Of course, there is still an on-going debate about which multi-factor model best describes the investment opportunity set. For our analysis, we relied on the most common models.

Green (2004) and Berk and van Binsbergen (2015) in that we cannot reject the hypothesis that it is a zero-sum game after costs.[9] Furthermore, we see periods in which the average active managers underperform the index alternatives, such as, before the dot-com bubble burst, during the financial crisis, or in the very recent period from 2014 to 2016. However, we can also observe periods in which active managers, on average, outperform, such as, from 2000 to 2007 or from 2009 to 2014.

Analyzing the drivers of the difference in the performance of active and index funds, we find that the equity and fixed income active managers have less exposure to traditional risk factors such as market and duration risk. Instead, active equity funds have a small cap and growth stock bias and active fixed income funds load on credit risk. Surprisingly, when the market is affected by unexpected volatility shocks, active management tends to underperform the average index investor. In periods of calm markets and when the implied volatility decreases, active managers tend to outperform. We explain this finding as being due to active managers who prefer to sell insurance and generate exposures to risk premia that perform well in good times but may cause substantial losses in bad times. The significant higher exposure to small cap companies for equity and credit risk for fixed income managers supports this hypothesis.[10]

Our data also allows us to shed light on the difference between retail and institutional funds. We find that after fees, there are a majority of unskilled mutual funds for the retail segment. In contrast, we see a more balanced proportion of skilled and unskilled funds for the institutional sectors and outside of the US. Thus, our results provide direct evidence for Garleanu and Pedersen (2018), who argue that more sophisticated investors outperform small investors because of the higher economies of scale in searching for skilled active managers.[11] Moreover, our results endorse the hypothesis of Gennaioli et al. (2015), who claim that trust is an essential component of the high fees in asset

[9] Furthermore, our findings resonate well with Pastor et al. (2015), who argue that in markets in which the mutual fund industry is big, such as the "US Equity Large Cap Blend," active alphas tend to be negative, and the equal-weighted alpha within investment categories exceeds the value-weighted alpha.

[10] Agarwal and Naik (2004) find similar return patterns for hedge funds. Thus, mutual funds try to profit from the same opportunities as hedge funds but have of course a narrower set of investment opportunities, due to regulatory restrictions. However, our results seem to contradict some of the previous findings, such as those of Moskowitz (2000) and Kosowski (2011), among others. They find that actively managed mutual funds tend to perform better than their passive benchmarks in bad times. However, these papers do not cover the recent financial crisis.

[11] Also, our empirical analysis supports the findings of Gerakos et al. (2016), who show that institutional investors outperform their strategy benchmarks after fees. Their analysis is based on self-reported but GIPS (Global Investment Performance Standard) compliant data, which still may inherit some biases, while our analysis is based on publicly available performance data.

management, and who argue that active retail managers profit from pandering to trusting investors by buying hot assets, which explains the tendency for active retail mutual funds to have positive exposure to growth stocks.

We further demonstrate, along the lines of Carhart (1997), that the average active retail investor can significantly improve their performance over the period ranging from 1993 to 2016, provided the worst-performing active mutual funds of the past year are neglected. However, when the investor concentrates only on the top performing funds, the overall performance cannot be significantly improved. In addition, we explore the role of fund size and fees on fund performance. Sorting active fund portfolios according to their performance persistence, fees, and size, we find that winner portfolios with low-fee and small funds tend to outperform but their alpha does not survive our test statistics. However, for both equity and fixed income retail funds, we find that a fund investor is well advised to avoid high-fee and small losers, as they generate significantly negative alphas.

The remainder of this paper is organized as follows. Section 2 discusses the data and performs a preliminary analysis, which motivates the design of our empirical tests. Section 3 presents our robust alpha test and the multiple hypothesis framework. In Section 4, we first compare the performance of single index and active mutual funds when benchmarked against factor models and an investable index. In Section 5, we provide a comparison of the value-weighted performance of active and index mutual funds portfolios across investment categories and asset classes. We analyze the drivers of the difference in the performance of active and index funds, and we explore the role of performance persistence, fund size, and fees. Section 6 concludes.[12]

2 Preliminary analysis

After describing our data, we analyze the potential time and cross-sectional dependencies in mutual fund returns to guide the formulation of appropriate test statistics for our hypotheses.

[12] All Matlab code used in this paper is available from the authors on request.

2.1 Data

Our mutual fund sample is drawn from the Morningstar database and ranges from December 1991 to December 2016. We include a total of 61,269 funds from different asset classes.[13] Table 1 shows the summary statistics of cross-sectional monthly attributes across asset classes. For the active funds, we analyze in all 14,969 institutional and 46,300 retail funds, while we have 56,136 active and 5,133 index funds. In general, there are fewer index funds, but they show higher average total net assets (TNA) and net returns, and also lower fees and about the same average years in the database. As expected, the institutional funds charge lower fees than their retail counterparts.

[Table 1 about here.]

As of December 2016, the total net assets of equity retail funds amounted to USD 9 trillion, those of fixed income retail funds to USD 3.7 trillion, and those of equity institutional funds and fixed income institutional fund to USD 3.1 trillion and USD 2 trillion, respectively. Since institutional investors often invest their money through mandates, there are fewer institutional funds than retail funds. The assets under management for index funds have been steadily increasing over our sample period. By the end of 2016, we find the highest concentration of index funds for equity funds, with 28% for retail and 32% for institutional funds. Looking at the fixed income funds, we find 18% of the retail and 13% of the institutional funds were index funds. For a more detailed description of the data and the data cleaning procedures, we refer to Appendix A.

2.2 Dependency analysis

It is well known that statistical inference for econometric models is severely complicated by the existence of serial and cross-sectional dependencies. Fama and French (2010) find that cross-sectional dependence can materially change the inference and, therefore, propose an appropriate adjustment for their single fund analysis.[14] At the same time, they correctly point to a potential caveat in

[13] In comparison, Pastor et al. (2015) explore 3,126 actively managed US equity-only mutual funds while Berk and van Binsbergen (2015) use 5,974 actively managed funds. Hence, we add to the existing literature by providing evidence based on our new dataset. Furthermore, to the best of our knowledge, we are the first to apply a robust multiple hypothesis framework to active and index mutual funds in an international context.

[14] See also Chen et al. (2017).

their resampling approach. Because they perform a random sampling of months, they lose any effects of autocorrelation. Similarly, neither does Barras et al. (2010) take into account serial dependence, claiming that they find such an effect only for a few mutual funds.[15] Moreover, while high cross-sectional dependencies could potentially bias their estimators, they find a low average pairwise correlation of 0.08 in their sample and argue that the cross-sectional dependencies are sufficiently low to allow consistent estimators. Given that we analyze not only the returns of single funds as in Barras et al. (2010) and Fama and French (2010) but also of mutual fund portfolios, and compare them against multi-factor and investable benchmark models, we must test for time dependence in a variety of settings. Therefore, we take a closer look at our data.

To compare active and index investing, we construct two types of different benchmark models. First, we apply the commonly used multi-factor models. In particular, for the equity analysis, we use the regional five-factor model with "market," "size," and "value" factors as given in Fama and French (1992) and add the "momentum" factor of Jegadeesh and Titman (1993) as well as the "betting against beta" factor of Frazzini and Pedersen (2014).[16] For the fixed income analysis, we apply a four-factor regional model including the "shift," "twist" and "butterfly" factors, as well as MSCI Inc.'s credit risk factor, measured as the BBB–AAA spread.[17] Second, we focus on the investable one-factor benchmark model, which we build as the value-weighted return of the index funds within the investment category of the analyzed time-series.[18]

We first test for serial dependence, applying the classical Ljung–Box (LJ) test and, as a robustness check, the distribution-free test of Genest and Rémillard (2004).[19] For both tests we must fix the number of lags L, for which we use the automatic block-length selection for the dependent bootstrap of Politis and White (2004) and the correction of Patton et al. (2009). We find that most mutual funds show an optimal block size of two or three. Therefore, we set the lag L to three for the two

[15] However, recently, Zhang and Yan (2018) find that the standard bootstrap can be misleading.

[16] The regional factors were retrieved from the homepage of Kenneth French, while the "betting against beta" factor is provided for each region on the homepage of AQR.

[17] The factors "shift," "twist," and "butterfly" represent the risk for a change in the level, steepness, and curvature of the term structure. See DeMond et al. (2012) for a description of the factors.

[18] For the investable one-factor model, we require at least 12 monthly returns and for the multi-factor models at least 36 monthly returns.

[19] Much criticism has been leveled at the possible low power of the LJ test. The LJ test is based on autocorrelations and, hence, it is not a real test of independence. The test developed by Genest and Rémillard (2004) uses ranks and, therefore, is distribution-free and does not depend on the underlying distribution of the observations.

tests.

[Table 2 about here.]

Table 2, Panel A, presents the number and percentage of funds that have a p-value below 5% based on the standard LJ test and the test of Genest and Rémillard (2004). Both tests provide us with a similar pattern. We find that for single equity funds, the percentage of rejected null hypotheses of no serial dependence over time ranges from 14% to 23%, with a slightly higher rejection rate for retail funds. Similarly, the percentage of fixed income funds rejecting the null ranges between 12% and 22%. For fund portfolios, the rejection rates can even be as large as 80%, although the number of portfolios is rather small and, hence, the results have to be interpreted with care. Nevertheless, it becomes clear that single fund and portfolio residuals are not serially independent. Furthermore, when we go into more detail, we find that the single mutual funds with the longest available time-series show a higher percentage of rejections. For example, the 2% oldest equity and fixed income single mutual funds, benchmarked against the investable model, exhibit statistically significant serial dependence in 40% and 63% of the cases.[20] Overall, signs of serial dependence can be found in roughly every fifth single mutual fund, and every third mutual fund portfolio. This evidence clearly justifies the need to control for dependence over time when we analyze the alpha of single and portfolios of mutual funds against different benchmark models.

We next test for cross-sectional dependence, which might occur if mutual funds "herd" in their holdings, as is shown by Wermers (1999). To detect cross-sectional dependence in our data, we apply the test of Pesaran (2004).[21] To compute the test statistic, we concentrate on funds that have more than one time period in common. Panel B of Table 2 presents the average pairwise correlation of the residuals together with the p-values of the Pesaran (2004) test. We find that we can reject the null hypothesis of no cross-sectional dependence at the one percent significance level for all single and portfolio fund categories.

When we compute the average pairwise correlation in our single fund sample, we find the same

[20] We do not report these more detailed results here, but they can be obtained from the authors.

[21] Compared to the well-known Lagrange multiplier test of Breusch and Pagan (1980), this test is correctly centered for a large sample and comparably short time-series, which is precisely the case here, since we have a broad cross-section of mutual funds but a comparably small time-series. Also, Pesaran (2004) finds that the test has satisfactory power even under weak cross-sectional dependence.

value for our US equity fund sample as Barras et al. (2010), i.e., an average of 0.08. For the single mutual funds with the investable benchmark models, we find average correlations between 0.04 for retail equity and 0.15 for institutional fixed income funds. For the multi-factor benchmarks, the correlation turns out to be considerably higher: 0.22 for retail equity funds and 0.62 for retail fixed income funds. On the portfolio level, we find average pairwise correlations of 0.13, 0.09, and 0.43 for the investment categories, equity, and the fixed income mutual fund portfolio, respectively. Although Barras et al. (2010) are not overly concerned with cross-sectional dependencies, we cannot merely use their reasoning given the elevated levels of average pairwise correlation, especially for fixed income funds.

3 Robust alpha test and multiple hypotheses

The above empirical evidence dictates that statistical tests must take into account both serial and cross-sectional dependence. To control for serial dependence, we propose a robust alpha test based on a studentized block bootstrap, which improves the accuracy of an inference for dependent time-series data compared to other methods.[22] To compute the bootstrapped t-statistics and p-values we closely follow Ledoit and Wolf (2008, 2011), who study the related problem of testing whether two Sharpe ratios or two variances are equal. We outline the mathematical details of the bootstrapped standard error of the estimated alpha in Appendix B. Once we have calculated the bootstrapped t-statistics and p-values in Equations (B.12) and (B.13), we can use them as input for multiple hypothesis testing.

While we control for serial dependence for the single-hypothesis alpha test, we control for cross-sectional dependence for the multiple-hypothesis method. Depending on whether we analyze single funds or portfolios of funds, we control either the false discovery rate (FDR) or the family-wise error rate (FWER).[23] As Bajgrowicz and Scaillet (2012) argue, investors do not rely on a single active manager but instead diversify between different managers. Therefore, in their view, it is favorable to control the amount of falsely rejected hypotheses (FDR) instead of investing only in the best

[22] See, e.g., Lahiri (2003), Haerdle et al. (2003), and Ledoit and Wolf (2008, 2011).

[23] The FWER dates back to Bonferroni (1936), and is defined as the probability of at least one false discovery. Romano and Wolf (2005a,b) introduce a stepwise multiple testing procedure that not only has higher statistical power than the tests of Bonferroni (1936) and Holm (1979) but also allows for cross-sectional dependence. The FDR is defined as the expectation of the proportion of falsely rejected null hypotheses. For a larger number of hypotheses, Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) show that it is favorable to control for the FDR.

strategies, as is the case when controlling the more conservative FWER. This diversification argument of Bajgrowicz and Scaillet (2012) no longer holds for portfolios of funds. Therefore, we control for the FDR when analyzing single funds and for the FWER when analyzing portfolios of funds.

For the FDR, we rely on the approach of Barras et al. (2010). A strength of their approach is that it regards each fund in isolation. However, this advantage comes at the cost that a high cross-sectional dependence could potentially bias their estimators. Thus, we conduct a set of Monte Carlo simulations for the multi-factor model in the fixed income market, where the findings in Panel B of Table 2 have indicated a high degree of cross-sectional dependence. In unreported results, we find that the average estimates are less stable but still close to the estimates where we assume independent residuals.[24] Therefore, we are confident that, for our application, we have consistent estimators also for the markets with a higher degree of herding. However, we must be aware that for the single fixed income market outside of the US, we have a higher estimation error.

While for the analysis of single funds, we have a large number of hypotheses, we only have a few hypotheses for the analysis of portfolios of mutual funds. Hence, we prefer to control for the FWER, applying the state of the art multiple hypothesis framework of Romano and Wolf (2016), which provides an efficient way to calculate the adjusted p-values. Since for the fund portfolios we have no missing values or disconnected time-series, as is the case for single funds, we can jointly sample blocks of fund and benchmark returns, thereby taking fully into account cross-sectional dependence. As for the FDR, we sample the test statistics with our robust alpha test, which allows us to take into account the serial dependence structure.

[Table 3 about here.]

Table 3 summarizes the motivation for our estimation strategy. Our test based on the block bootstrapped alpha is, in combination with the FDR (for single funds) and the FWER (for fund portfolios), a suitable method taking into account both serial and cross-sectional dependence simultaneously, as evidenced by our preliminary analysis. Also, both frameworks take into account the characteristics of the data. For the single fund analysis with a large cross-section and a small overlap of the time-series, we regard each fund in isolation and therefore prefer to control the FDR. For the

[24] These results can be obtained from the authors.

portfolio analysis with a small cross-section and fully connected time-series, we focus on the more restrictive FWER and jointly block bootstrap the entire data sample. To explore the accuracy of our test, we present the results of a simulation exercise in Appendix C. We find that our robust alpha test is still liberal but more accurate since it also corrects for the serial dependence observed in the data. The standard inference tests are too liberal in rejecting the null hypothesis. Thus, when we apply the standard tests or sample only one return each time instead of a block of returns, we generate more type I errors (false positive findings) than expected by the test.[25]

4 Single mutual funds

We first analyze the distribution of the single fund alphas measured against the regional multi-factor benchmark models. While such a comparison is not suited to identifying skilled managers, it gives us an idea about the risk drivers and style exposures of the different funds. Later, we compare the single funds' alphas benchmarked against an investable index.

4.1 Multi-factor benchmark

For equity funds, the five-factor model is used as the multi-factor benchmark. It is based on the three-factor model of Fama and French (1992) but includes the "momentum" and "betting against beta" factors. For fixed income funds, we rely on MSCI's four-factor model with the corresponding regional factor returns. As the first step, we calculate our robust alpha, where for each individual fund, we apply the optimal block size with the method of Politis and White (2004) and Patton et al. (2009).[26] For the multiple-hypothesis adjustment, we control for the FDR using Barras et al. (2010) and compute the proportion of negative, zero, and positive alphas after fees. It is important to emphasize that we equal-weight each fund, since the method of Barras et al. (2010) regards each fund in isolation and does not allow value-weighted adjustments. Table 4 reports the results.

[Table 4 about here.]

[25] We also note that even if there is no serial dependence, our block-bootstrapped alpha test statistic is accurate.
[26] For robustness, we also applied a block size of six, which yields the same results.

For equity funds, we find that the proportion of active funds with zero alpha is 62.3% for retail and 77.4% for institutional funds. As expected, the proportion of zero alpha funds for index investors is higher, at 68.4% for retail and 79.5% for institutional funds. We also find that the percentage of funds with a significantly positive alpha is the highest for active institutional funds, 3.5%, while for all the other categories there are between 1.4% and 1.9% single mutual funds with a positive alpha. The proportion of funds with a negative alpha is the highest for active retail funds, 35.9%, while active institutional funds have 19.1% of their managers generating a negative alpha. Surprisingly, we also find that 29.7% of the retail and 19.1% of the institutional index funds provide a negative alpha. At the same time, we observe 19.1% of the institutional active funds with a negative alpha. Hence, for institutional funds, we put both active and index managers at a similar disadvantage by using a multi-factor benchmark if we were to interpret the resulting alpha as skill. Moreover, for the US market, there are more institutional index funds (33.1%) than institutional active funds (30.7%) having a negative alpha.

Focusing on US institutional active equity funds, we find that our 69.3% zero funds compares well with the 75.4% of Barras et al. (2010, Table II). At the same time, we have a higher fraction of negative alpha funds, 30.7% against their 24%. Considering the impact of luck in the left tail, however, the proportion of significantly negative alphas (*FDR 10 alpha* < 0) drops to 3%, which is considerably smaller than their 13.6%. Interestingly, the proportion of significantly negative alphas for index funds only drops to 17.2%. From this perspective, index funds seem to perform even worse than active funds, when benchmarked against a multi-factor model.

For fixed income funds, we observe that for retail active funds, there is an equal number of negative and positive alphas (around 16%), while for retail index funds we find a smaller fraction of negative alpha funds in favor of a higher fraction of zero alpha funds. For institutional funds, we find more positive than negative alpha funds. Hence, as with equity funds, retail funds seem to perform worse than institutional funds if benchmarked against multi-factor models. For the regions outside the US, we do not find active funds in the fixed income universe with a negative alpha. For the US, the shares of negative, zero, and positive alphas are of similar magnitudes.

While these comparisons are informative about the style and risk exposure of the different funds, they neither represent manager skill since the alphas are not value-weighted, nor do they provide

useful information for return-chasing fund investors since multi-factor models provide only an unfair benchmark.[27] The inappropriateness of multi-factor benchmarks for performance measurement becomes most evident from the observed negative alphas of the index funds. If multi-factor benchmarks were fair, then index funds should have zero alpha on average.

4.2 Investable benchmark

Given the above concerns, we next construct investable benchmarks based on Morningstar's investment categories. We rely on these categories as they are well established in the industry, and their definition perfectly serves our intention to benchmark active funds.[28] To construct an investable benchmark, we value-weight all index funds within a given category. By value-weighting the index funds in each category, we obtain the investable benchmarks which we use for calculating the alphas of active funds. To make the analysis comparable to the previous section, we only include those Morningstar's investment categories that include the investment regions US, Global, Europe, Japan, and Asia ex-Japan for the equity funds. For the fixed income market, we include the categories in US dollar US, Swiss Franc CHF, Europe EU, and Sterling GBP. We first calculate the p-values from our robust alpha test, after fees and with active funds benchmarked against the corresponding value-weighted category index. We then compute the estimated percentage of negative, zero, and positive alpha funds using the method developed by Barras et al. (2010).

[Table 5 about here.]

Table 5 shows the results using the same categorization as in Table 4. Strikingly, we find much higher averages of zero alpha active funds for both retail and institutional funds. In particular, for the US and the Global categories, the difference is substantial. For instance, with an investable index as benchmark, the fraction of zero alpha active US institutional funds rises from 69.3% to 80.1% and the proportion of significant negative alphas decreases from 25.3% to 0.2%. At the same time, the

[27] See, Berk and van Binsbergen (2015).

[28] We acknowledge that there are many routes to take for benchmarking fund portfolios. In practice, when investors or active managers focus on a specific investment category, they do not compare themselves with the multi-factor models in general, rather, they compare themselves with other funds within the same category. As Morningstar states on its website, "the classifications were introduced in 1996 to help investors make meaningful comparisons between mutual funds." While the investment objective stated in a fund's prospectus does not always reflect how the fund actually invests, Morningstar places funds in a given category based on their portfolio statistics and securities holdings.

proportion of significantly positive alpha funds remains at 0.0%. Hence, the large fractions of zero alpha funds after fees support the equilibrium argument of Berk and Green (2004).[29]

What surprises us in Table 5 is the remarkably large fraction of index funds with a negative alpha, especially in the US. The average of zero alpha index funds is below the average of zero alpha active funds. This observation may be due to the fact that, so far, we have ignored fund size in our analysis, since the framework of Barras et al. (2010) does not allow us to value-weight the funds's alphas. Fund size, however, is a crucial element in the argumentation of Berk and Green (2004). Therefore, we now analyze the funds' performance while taking into account fund size.

5 Value-weighted portfolios of mutual funds

As Berk and Green (2004) argue, funds managed by skilled managers attract greater portfolio flows than funds managed by unskilled managers. Hence, if we want to measure the skill of a fund manager, or if we want to test whether active investing is a zero-sum game, we must measure performance on a value-weighted basis and against an investable benchmark.[30]

5.1 Investable benchmark

For the performance analysis of fund portfolios, we focus again on the same investable benchmarks as we used in Section 4.2. Since we require connected time-series for our multiple hypothesis adjustment, we focus on the periods from 1993 to 2016 and 2000 to 2016, which allows us to include more investment categories for the more recent time periods. Given that index mutual funds only emerged recently, we observe for the period starting in 1993 at least one index fund for four institutional and 17 retail categories. For the more recent period starting in 2000, we obtain 30 investment categories

[29] Also, in Table 4, the proportion of active fixed income funds with significantly positive alphas is surprisingly high. With the investable index as benchmark, these numbers turn out to be much more moderate, pointing at the potential problem of defining appropriate multi-factor benchmarks. Hence, switching to an investable benchmark allows a much more realistic assessment of actively managed funds.

[30] As an additional exercise, we also benchmarked our portfolios of funds against multi-factor models using the same categorization as in the previous section. For active institutional US equity funds, our results show a significantly negative alpha and, therefore, are in line with Fama and French (2010). However, except for active institutional Global funds, all other active alphas are insignificant. In contrast, we find again some significantly negative alphas for index funds. For fixed income funds, we find for active institutional funds a significantly positive alpha. Since we have argued above that multi-factor benchmarks are not suitable for performance measurement, we do not report these results here.

for the retail segment and 12 investment categories for the institutional segment. Hence, we end up with 63 categories. By value-weighting the index funds in each category, we obtain the investable benchmarks which we use for calculating the alphas of active funds.

[Figure 1 about here.]

In Figure 1, we plot the robust p-values against the net and gross alphas for each of the available investment categories. As argued in Section 3, we adjust the p-values for multiple hypothesis testing using the method of Romano and Wolf (2016). After fees, we find the "US Equity Large Cap Blend" category for institutional funds and the "Canada Fixed Income" category for retail funds to significantly underperform the alternative of the value-weighted index funds for both periods. For the negative alpha of the "Euro Fixed Income" retail category and the period from 1992 to 2016 we also find a significant p-value. Hence, only for three investment categories can we reject the zero-sum game hypothesis of Berk and Green (2004). Furthermore, our finding that "US Equity Large Cap Blend" institutional funds underperform after fees is perfectly in line with the argument of Pastor et al. (2015), in that higher competition in big active mutual fund industries leads to diminishing returns to scale. Before fees, there are no investment categories with significantly negative alphas. However, we find "US Fixed Income" from 1992 to 2016, and "Global Equity Large Cap," "Emerging Markets Equity," and "Europe Equity Large Cap" from 2000 to 2016 for institutional, and also "Global Equity Large Cap" from 1992 to 2016 and "Global Equity" from 2000 to 2016 for retail funds to significantly outperform the value-weighted index funds.[31]

In Figure 2, we show the cumulated aggregated alpha of active index funds over time.[32] Equal weighted, all investment categories for equity funds provide a positive alpha over time, even after fees. Value weighted, the aggregated alphas remain positive before fees, but they are zero for institutional funds and slightly negative for retail funds after fees. For the fixed income mutual funds, we find that the value-weighted alpha across investment regions is positive for the equal and value-weighted

[31] We remark that the choice of a block size of three is a conservative choice. As a robustness check, when we apply a block size of six or nine, the p-values increase slightly. The "Euro Fixed Income," "Emerging Markets Equity," and "Europe Equity Large Cap" categories, which all exhibit a p-value just below 10% for the block size of three, start to show insignificant p-values, further supporting the theory of a zero-sum game after fees.

[32] We first compute the value-weighted alpha within an investment category against the value-weighted benchmark of the index funds and then aggregate the investment categories with equal and value-weights. We also split into retail and institutional funds before and after fees. For the equity mutual funds, we find that active mutual funds provide a superior alpha than index funds in every analysis before fees.

aggregation of the investment categories. For the institutional funds after fees, we also observe positive alphas over time. There are three major periods where active managers underperformed their index counterpart: equity funds before the burst of the dot-com bubble, both equity and fixed income funds in the financial crisis, and a slight underperformance in the recent past, especially after fees.

[Figure 2 about here.]

The fact that the equal-weighted alpha across investment regions is higher than the value-weighted alpha adds evidence to the theory of Pastor et al. (2015) in the sense that the higher competition in big active mutual fund industries decrease the fund's ability to outperform passive benchmarks. For the fixed income mutual funds, however, we observe that both weightings lead to roughly the same alpha over time. This pattern could indicate that the competition in the fixed income segment remains low, regardless of fund size, due to the higher complexity of the product. In addition, Garleanu and Pedersen (2018) argue that small investors tend to underperform because of their higher search costs and fees, while large investors are expected to outperform after a certain size because of their economies of scale and lower fees. For equity funds, we can confirm their theory. We observe higher aggregated alphas for institutional funds after costs, but similar alphas for retail equity managers before costs. Since retail funds can pool the investments of small investors, and mutual funds often manage retail and institutional money in the same aggregated fund, we expect this pattern. However, for fixed income funds, we find that retail funds achieve a much lower alpha after fees than their institutional competitors.

What is surprising in Figure 2, however, is the existence of three major periods where active managers underperformed their index counterparts: before the burst of the dot-com bubble, in the financial crisis, and in the recent past. Since we would have expected that active management pays in turbulent times, we will further explore this observation in the next section.

5.2 Drivers for the difference in performance of active and index funds

To gain further intuition about what drives a wedge between the performance of the average active fund and that of the average index fund, we ask whether the multi-factor model of Section 4.1 provides

some explanation for the difference in returns between the value-weighted portfolios of active and index mutual funds. Alerted by our observation from Figure 2, we enrich our regressions with the volatility index (VIX) of the Chicago Board Options Exchange (CBOE) as a fear gauge to proxy for market uncertainty.[33]

Table 6 shows the results for both active equity and fixed income funds when measured against the investable benchmark. Overall, the average values of R^2 absorb a significant fraction of the variance of the alpha before fees, in particular for fixed income funds and US institutional equity funds. For the equity funds in Panel A, we observe that the gross alpha loads profoundly and significantly on the SMB factor. Also, especially in the US, the performance difference loads negatively on the HML factor. The exposure is more pronounced for retail funds. Institutional funds, in contrast, have a much lower exposure to growth stocks. Overall, active funds seem to have a prominent small-cap bias and favor growth over value stocks. Furthermore, they tend to load positively on the momentum factor and negatively on the betting-against-beta factor.

[Table 6 about here.]

Concerning the VIX, we find that the difference in performance of active and index investing shows a negative sensitivity to changes in the VIX, which is often statistically significant. At first sight, this finding seems to run against our intuition, as we would expect active managers to use their skill to anticipate sudden uncertainty shocks. However, active managers that protect their portfolio against adverse shocks must pay an insurance premium in the long term. Such protection would generate relative losses to the market return in good times. Therefore, our result suggests that active managers prefer to run a short exposure to general market volatility, i.e., they tend to prefer small gains by selling insurance.

In Table 6, Panel B, we see that fixed income managers have a negative exposure to the shift factor. Consequently, they are less affected by rising interest rates. In exchange, they load on other risk factors to compensate for the lower expected returns. In particular, they load significantly on

[33] We downloaded the time-series of the VIX index of the Chicago Board Options Exchange (CBOE) from Bloomberg. In unreported results, we find that a high level of proxied uncertainty, e.g., by earnings-per-share volatility or dispersion of returns within a fund category, is in general favorable for the performance of actively managed funds. However, these effects are not significant.

the credit risk factor. As in Panel A for equity funds, fixed income managers also have a negative exposure to changes in the VIX. Therefore, they lose money if the VIX increases sharply, as it did during the latest financial crises, which also explains the large drop in the cumulative alpha in Figure 2 towards the end of 2008.

5.3 Persistence analysis

From an investor's perspective, it may be disappointing that active fund investing is, by and large, a zero-sum game after costs. How then can a fund investor do better and profit from actively managed funds? An initial idea is provided by Carhart (1997). He finds that US equity mutual funds with a substantial underperformance over the past year persist to underperform over the next year relative to a multi-factor benchmark. In contrast to the outperformance of the best mutual funds, he cannot explain the persistence in the worst mutual funds.[34] Thus, it would be of interest to know whether a fund investor is better-off if avoiding the losers of the past year.

To simulate the returns to an average active investor who trades according to this simple rule, we build momentum portfolios of active funds as follows. Every year in December, we first sort the active funds within each investment category based on their t-value for the value-weighted alpha measured against the investable benchmark.[35] Then, we invest in the value-weighted portfolio of the $x\%$ best performing active funds and normalize the weights each month. We repeat the same exercise for the $x\%$ worst performing funds. If one month there is no data for a particular fund, it disappears from the portfolio. To aggregate the performance numbers of the different investment categories, we value-weight the net returns by the total active assets.[36] We assume that funds do not charge transaction costs for incoming and outcoming investors.

[Table 7 about here.]

[34] Among others, Huij and Derwall (2008) find persistence in the US fixed income mutual funds market.

[35] To compute the alpha, we require at least ten of the twelve most recent monthly returns. When we sort by the t-value for the alpha, we consider the market risk of the fund and look at both the relative performance and the consistency of the relative performance against the benchmark.

[36] For $x\%$ we chose steps of 10% starting with all mutual funds to the best 10% mutual funds. We disregard data points where we have less than ten active mutual funds, and to calculate the benchmark return for the alpha we must have at least one index fund within the category for the past twelve months to start the out-of-sample backtest. For some small investment categories, over the year the number of funds drops below ten. In this case, we apply the next less restrictive filter.

Table 7 presents the momentum portfolio returns when selecting the best performing (Panel A) and the worst performing funds (Panel B). In Panel A, we find that the performance increases the more we focus on the best performing funds. For equity funds, the alpha after fees climbs from −0.23% (−0.60%) to a remarkable 0.62% (−0.02%) for institutional (retail) funds. For retail fixed income funds, alpha increases from −0.75% to −0.40%, but decreases for institutional fixed income funds from 0.26 to 0.16.[37] However, all these results are statistically insignificant, as the robust p-values remain high or even increase. Hence, after fees, even the best performing funds constitute a zero-sum game for the fund investor. This absence of persistence supports the theoretical argument of Berk and Green (2004) that persistence should not exist since new money flows into well-performing funds and there are diseconomies of scale, or because successful funds capture excess returns by raising fees.[38]

In Panel B of Table 7, we form portfolios by selecting the x% worst performing funds. We find that the value-weighted performance decreases drastically. For instance, for institutional equity funds it drops from -0.23% to -0.94%. However, only for retail funds does the negative performance of the 10% worst performing funds survive our robust alpha test adjusted for multiple hypothesis testing. For equity retail funds, the performance drops to -1.39% at the 5% significance level and, for fixed income retail funds, it drops to -0.84% at the 10% significance level. Hence, while Carhart (1997) shows for US equity funds and under a multi-factor benchmark that the persistence is significant for the worst performing funds, we can confirm this result only for retail equity and fixed income funds. Investing in these funds, obviously, is not a zero-sum game after costs. For institutional funds, we do not find such evidence. At the same time, Table 7 confirms Carhart (1997) in that we do not find any unexplainable persistence in overperforming funds.

Chen et al. (2004) and Yan (2008) find a negative relation between alpha and size and a positive relation with past return. Thus, on average, they find that future alpha is smaller for large funds but past returns are associated with higher future alpha, and predictability exists. Recently, Elton

[37] Interestingly, for fixed income institutional funds, we observe substantially lower betas the more we exclude badly performing funds from the portfolio, indicating that the best performing fixed income funds run a slightly different exposure than their investable benchmark suggests.

[38] As an additional exercise, we also explored the persistence of gross alphas. In unreported results, we find that the alphas increase substantially, but still they do not survive our statistical test, not even for the 10% best performing funds.

et al. (2012) show that alpha persistence does not disappear for larger funds. To explore the interplay between size and predictability, we perform a bivariate sort on size and persistence.

[Table 8 about here.]

Table 8 presents the alphas after fees for the different portfolios sorted according to fund size and the previous year's performance. We find that, except for fixed income retail funds, small winner funds perform better than their larger counterparts. They produce the highest alphas relative to their investable benchmarks. However, this outperformance is statistically insignificant. At the same time, we identify the small loser funds as the funds with the worst performance. For retail equity and fixed income funds, the negative alphas of small (and medium) loser funds become even statistically significant. Unsuccessful and small funds will continue to be unsuccessful, and they do so in a statistically significant way, while large funds tend to underperform but not significantly so. These findings resonate well with Berk and Green (2004)'s hypothesis that fund performance is inversely related to size due to diseconomies of scale.

5.4 Impact of fees

For US equity funds, Gil-Bazo and Ruiz-Verdú (2009) find a puzzling underperformance of mutual funds that charge higher fees. Their finding contradicts the argument of Habib and Johnsen (2016) that higher fees act as a signal for the unobservable quality of the costly research by active managers. However, higher fees can also be seen as a sure loss for investors, since they directly reduce the portfolio return when the quality of the manager is unobservable. Thus, higher fees imply lower net returns if the costly research of the active manager does not improve performance. To shed more light on this debate, we explore the impact of fees by proceeding analogously to the persistence analysis of the previous section.

[Table 9 about here.]

Using an investable benchmark, Table 9 presents the performance of active fund portfolios that include the x% least expensive (Panel A) and the x% most expensive (Panel B) funds of the preceding

year. In Panel A, the alpha against the investable benchmark increases for institutional equity funds from −0.23% to 0.83% and from −0.60% (−0.75%) to 0.30% (−0.40%) for retail equity (fixed income) funds. Hence, even the alpha of retail equity funds gets into positive territory if we exclude those funds that charge the highest fees. Again, although a fund investor can improve the performance in their fund portfolio by including only the least expensive funds, these improvements are statistically insignificant. For institutional fixed income funds, the alpha first increases but then decreases for the fund portfolio with the 20% and 10% lowest fees. Interestingly, the betas of these portfolios are substantially below one.

Panel B of Table 9 shows the performance of the portfolios with the x% most expensive funds. If the argument of Habib and Johnsen (2016) were valid and high fees were a signal of quality, we would expect increasing alphas the more we filter out the cheaper funds. However, we observe the opposite. The portfolios with the 10% most expensive funds perform poorly. For instance, the performance of the portfolio of the 10% institutional equity funds drops to –1.14%, compared to the 10% cheapest fund with an alpha of 0.83%. For equity retail funds, the underperformance of high-fee funds becomes significant already when we look at the 90% most expensive funds. For the 10% most expensive retail funds, we get a highly significant alpha of –2.83%, compared to the 10% least expensive retail funds with an alpha of 0.30%. Hence, if we only consider the universe of the most expensive equity retail funds, active investing definitely is no zero-sum game.

[Table 10 about here.]

Given the evidence that past winners and low fee funds generate a higher average alpha over time, we next ask whether the same pattern emerges when we control for performance persistence and fees simultaneously. Hence, we build nine portfolios that arise from the bivariate sort and the 30th and 70th percentiles for each criterion.[39] Table 10 shows the results. All active alphas are positive for the low-fee and winner portfolios except for the fixed income retail investor. Furthermore, for all sorted portfolios, the alpha in the top right (low fee and winner) corner is always larger than the alpha in the lower left (high fee and loser) corner. These alphas are insignificant, except for the equity retail funds. Here, the high-fee and loser portfolio has a highly significant negative alpha. Overall, we find

[39]For some small investment categories, there are time periods where none of the mutual funds belong to a particular group. In such a case, we invest in the value-weighted portfolio of all active funds within this category.

that both high performance and low fees over the past year have a positive impact on the alpha in the next year. The result is robust in the bivariate sort of the two criteria. Especially for retail funds, an investor is well advised to avoid high-fee loser funds as the negative alphas are highly significant under our testing framework.

[Figure 3 about here.]

To shed further light on which funds charge higher fees, Figure 3 shows the average active fees of the highly competitive US equity market over time together with the relative share of index funds in terms of assets under management. As expected, we find a substantial difference between the fees charged by retail and institutional funds, depending on the fund's age. Young retail funds charge the highest fees. However, the size of their fees has drastically decreased since the recent financial crisis, converging to the level of the fees charged by older retail funds. Interestingly, over the whole period, young institutional funds have charged fees similar to their older competitors. The gap between old retail and old institutional funds has been somewhat steady over the years, slightly narrowing recently. We also find that expense ratios are lower for large funds.[40] As Figure 3 suggests, over the years, the level of fees for active funds has tended to decrease further. Thus, we find further evidence for the zero-sum game hypothesis of Berk and Green (2004) in the sense that active managers start to adjust their fees due to the unabated growth of index funds.

5.5 How different are the results without the robust alpha test?

In our preliminary analysis, we demonstrated that fund returns are serially dependent. Furthermore, the simulation study in Appendix C indicated that our robust alpha test based block resampling is still liberal, but closer to the nominal level of the test. Consequently, we expect that the additional discoveries under the alternative test statistics are false positives. Still, it is natural to ask how our test, applied to our fund data, compares to other tests if we ignore serial dependence, and whether it would make a material difference to our empirical results.

[40] By the end of 2016, we find that the average expense ratio of the equal-weighted portfolio compared to the value-weighted portfolio is 21% higher for institutional equity funds, 37% higher for retail equity funds, 17% higher for institutional fixed-income funds, and 34% higher for retail fixed-income funds. Since institutional funds are usually larger than retail funds, this gap is consistent with the finding of Elton et al. (2012) that expense ratios are lower for larger funds.

We find that for the analysis of a single fund's performance compared to the investable benchmark, and in the single hypothesis setting, our robust alpha test decreases the proportion of unskilled (skilled) funds from 11.2% (5.4%) for the standard test to 9.0% (4.3%) at the 5% significance level. Hence, using the robust block-sampling test decreases the proportion of nonzero-alpha funds by almost 20%, compared to the standard test. For the fund portfolios, the decrease is almost 31%.

[Table 11 about here.]

For multiple hypothesis testing using FDR for single mutual funds, the impact of using our robust test is substantial. The decrease in nonzero-alpha funds when using our robust alpha test is 69% compared to the standard test and 60% compared to the standard resampling test at the 5% significance level. At the 10% significance level, the proportion of unskilled funds drops from 7% to 3.8% and from 1.6% to 0.8%, which corresponds to a decrease of nonzero-alpha funds of roughly 50%. Similarly, for the FWER applied to fund portfolios, we get a reduction of 27% when we use block resampling to take into account serial dependence. These differences are substantial and represent false discoveries due to the ignoring of serial dependence by the standard resampling. Admittedly, the proportions of skilled and unskilled funds are small. However, for an investor selecting a specific fund investment, discriminating between a skilled and unskilled fund or between fund portfolios with zero and nonzero alphas is absolutely crucial. Using our test statistics, we adequately take into account two prominent features of the data: serial and cross-sectional dependence. Thereby, we avoid potential distortions in the test statistics.

6 Conclusion

Analyzing a rich dataset from Morningstar covering 61,269 mutual funds from different regions and asset classes from 1992 to 2016 and comparing their returns to those from common multi-factor models, we find that a large fraction of active equity managers show zero alphas after fees. However, when we conduct a fair performance evaluation for equity and fixed income mutual funds, we find significant negative alphas after fees only for the "US Equity Large Cap Blend" for institutional funds and "Canada Fixed Income" for retail funds. For the vast majority of categories, we cannot reject

the hypothesis that active investing constitutes a zero-sum game after costs. Indeed, we even find categories such as "US Fixed Income" and "Global Equity Large Cap" for institutional investors with significant p-values before fees.

At first glance one would expect active managers to invest more carefully and take fewer risks. We have confirmed this hypothesis by the fact that active management takes a more conservative position with respect to the traditional risk factors, such as market and duration risk. However, we find that active equity and fixed income mutual funds are affected by adverse volatility shocks, suggesting that active managers sell protection in order to collect the insurance premium. Also, we find that, averaging over the different regions, the active investor has a higher sensitivity than index funds to alternative risk premia such as small cap and credit risk.

Sorting active fund portfolios according to their performance persistence, fees, and size, we find that low-fee winner portfolios and small winner portfolios tend to outperform but their alpha does not survive our test statistics. These results give further support to the zero-sum game argument of Berk and Green (2004). Our analysis also highlights some substantial differences between institutional and retail funds. In particular, our empirical results suggest to active retail investors that they should avoid avoid high-fee losers and small losers. Their alphas are negative and statistically significant, surviving our robust test statistics adjusted for multiple hypotheses.

References

Agarwal, Vikas, and Narayan Y. Naik, 2004, Risks and portfolio decisions involving hedge funds, *The Review of Financial Studies* 17, 63–98.

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical trading revisited: False discoveries, persistence tests, and transaction costs, *Journal of Financial Economics* 106, 473–491.

Banegas, Ayelen, Ben Gillen, Allan Timmermann, and Russ Wermers, 2013, The cross section of conditional mutual fund performance in European stock markets, *Journal of Financial Economics* 108, 699–726.

Barras, Laurent, Olivier Scalliet, and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance* 65, 179–216.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* 29, 1165–1188.

Berk, Jonathan B., and Richard C. Green, 2004, Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269–1295.

Berk, Jonathan B., and Jules H. van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 118, 1–20.

Berk, Jonathan B., and Jules H. van Binsbergen, 2017, Mutual funds in equilibrium, *Annual Review of Financial Economics* 9, 147–167.

Bonferroni, Carlo E., 1936, *Teoria statistica delle classi e calcolo delle probabilita* (Libreria Internazionale Seeber).

Breusch, Trevor S., and Adrian R. Pagan, 1980, The Lagrange multiplier test and its applications to model specification in econometrics, *Review of Economic Studies* 47, 239–253.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey D. Kubik, 2004, Does fund size erode mutual fund performance? The role of liquidity and organization, *American Economic Review* 94, 1276–1302.

Chen, Yong, Michael Cliff, and Haibei Zhao, 2017, Hedge funds: The good, the bad, and the lucky, *Journal of Financial and Quantitative Analysis* 52, 1081–1109.

DeMond, Andrew, Erdem Ultanir, and John Fox, 2012, Barra term structure models, Technical report, MSCI Inc.

Elton, Edwin J., 2001, A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases, *Journal of Finance* 56, 2415–2430.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 2012, Does mutual fund size matter? The relationship between size and performance, *The Review of Asset Pricing Studies* 2, 31–55.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance* 65, 1915–1947.

Ferreira, Miguel A., Aneel Keswani, António F. Miguel, and Sofia B. Ramos, 2013, The determinants of mutual fund performance: A cross-country study, *Review of Finance* 17, 483–525.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

French, Kenneth R., 2008, Presidential address: The cost of active investing, *Journal of Finance* 63, 1537–1573.

Garleanu, Nicolae B., and Lasse H. Pedersen, 2018, Efficiently inefficient markets for assets and asset management, *Journal of Finance* forthcoming.

Genest, Christian, and Bruno Rémillard, 2004, Test of independence and randomness based on the empirical copula process, *Test* 13, 335–369.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, 2015, Money doctors, *Journal of Finance* 70, 91–114.

Gerakos, Joseph, Juhani T. Linnainmaa, and Adair Morse, 2016, Asset managers: Institutional performance and smart betas, NBER Working Papers 22982, National Bureau of Economic Research.

Gil-Bazo, Javier, and Pablo Ruiz-Verdú, 2009, The relation between price and performance in the mutual fund industry, *The Journal of Finance* 64, 2153–2183.

Habib, Michel A, and D Bruce Johnsen, 2016, The quality-assuring role of mutual fund advisory fees, *International Review of Law and Economics* 46, 1–19.

Haerdle, Wolfgang, Joel Horowitz, and Jens-Peter Kreiss, 2003, Bootstrap methods for time series, *International Statistical Review* 71, 435–459.

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 65–70.

Huij, Joop, and Jeroen Derwall, 2008, "Hot hands" in bond funds, *Journal of Banking and Finance* 32, 559–572.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65.

Kosowski, Robert, 2011, Do mutual funds perform when it matters most to investors? US mutual fund performance and risk in recessions and expansions, *The Quarterly Journal of Finance* 1, 607–664.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* 61, 2551–2595.

Kuensch, Hans Rudolf, and Friedrich Goetze, 1996, Second-order correctness of the blockwise bootstrap for stationary observations, *The Annals of Statistics* 24, 1914–1933.

Lahiri, Soumendra N., 2003, *Resampling Methods for Dependent Data* (Springer–Verlag).

Ledoit, Oliver, and Michael Wolf, 2008, Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Ledoit, Olivier, and Michael Wolf, 2011, Robust performances hypothesis testing with the variance, *Wilmott* 2011, 86–89.

Moskowitz, Tobias J, 2000, Discussion, *The Journal of Finance* 55, 1695–1703.

Newey, Whitney K., and Kenneth D. West, 1994, Automatic lag selection in covariance matrix estimation, *The Review of Economic Studies* 61, 631–653.

Pastor, Lubos, Robert F. Stambaugh, and Lucian A. Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23 – 45.

Patton, Andrew, Dimitris N. Politis, and Halbert White, 2009, Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White, *Econometric Reviews* 28, 372–375.

Pesaran, Hashem M., 2004, General diagnostic tests for cross section dependence in panels, CESifo Working Paper Series 1229, CESifo Group, Munich, Germany.

Politis, Dimitris N., and Halbert White, 2004, Automatic block-length selection for the dependent bootstrap, *Econometric Reviews* 23, 53–70.

Ratcliff, John W., and David E. Metzener, 1988, Pattern-matching: the gestalt approach, *Dr Dobbs Journal* 13, 46.

Romano, Joseph P., and Michael Wolf, 2005a, Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* 100, 94–108.

Romano, Joseph P., and Michael Wolf, 2005b, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P., and Michael Wolf, 2016, Efficient computation of adjusted p-values for resampling-based stepdown multiple testing, *Statistics & Probability Letters* 113, 38–40.

Sharpe, William F., 1991, The arithmetic of active management, *Financial Analysts Journal* 47, 7–9.

Wermers, Russ, 1999, Mutual fund herding and the impact on stock prices, *The Journal of Finance* 54, 581–622.

Yan, Xuemin Sterling, 2008, Liquidity, investment style, and the relation between fund size and fund performance, *Journal of Financial and Quantitative Analysis* 43, 741–767.

Zhang, Huazhu, and Cheng Yan, 2018, A skeptical appraisal of the bootstrap approach in fund performance evaluation, *Financial Markets, Institutions & Instruments* 27, 49–86.

A Description of the data

We summarize the steps for the data cleaning of the Morningstar database and provide summary statistics for the different asset classes and investment categories.

A.1 Raw Morningstar data

Our mutual fund sample is from the Morningstar database.[41] We focus on all funds with an *Investment Type* flagged by "Open-End Fund" or "Exchange-Traded Fund" including non-survivors from December 1991 to December 2016. We downloaded the following fields for each share class.

For the description of a share class, we retrieved the *Name*, *ISIN*, and *Base Currency*. It is common to name a share class starting with the name of the asset manager, followed by a description of the strategy, and an ending for the share class. For example, for the equity fund "Blackrock S&P 500 Index," there is a share class "Blackrock S&P 500 Index Institutional" for institutional and the "Blackrock S&P 500 Index Investor A" for retail clients.

The most specific categorization in Morningstar is the *Morningstar Category*, which is derived by analyzing the underlying portfolio holdings. In all, we find 504 different groups for the retail equity and fixed income funds. The *Global Category* combines several Morningstar categories, and we see a total of 68 groups for retail equity and fixed income funds. For example, the *Global Category* category "Europe Equity Large Cap" includes Morningstar categories, such as "EAA Fund Europe Large-Cap Blend Equity," "EAA Fund Europe Large-Cap Value Equity," "EAA Fund Europe Large-Cap Growth Equity," but also "US Fund Europe Stock" or "Canada Fund European Equity." Since we have within this broader categorization a higher chance of finding both index and equity funds, we concentrate on the *Global Category*. The *Global Broad Category Group* further aggregates the *Global Category* into the major asset classes. Since we focus on the comparison of active and index funds, we concentrate on the *Global Broad Category Group* "Equity" and "Fixed Income" funds. We thereby

[41] Recent work in Kosowski et al. (2006); Fama and French (2010); Barras et al. (2010) concentrates mostly on the survivor-bias-free CRSP US Mutual Fund Database. As shown by Elton (2001), the CRSP database also suffers from a survivorship bias: the so-called omission bias. Berk and van Binsbergen (2015) find that neither the CRSP nor the Morningstar database are free from errors. Thus, we must be careful, and we find the same errors as reported in this previous paper.

disregard categories such as "Allocation," "Money Market," or "Commodities" because for them there are insufficiently many index funds to make a fair comparison.

For the computation of the returns, we downloaded the following fields for each fund: *Monthly Return USD*, *Monthly Gross Return USD*, and *Net Assets - share class (Monthly) USD*. The *Monthly Return USD* includes management, administrative, and other costs that are deducted from the NAV, such as the 12b-1 fee. All income and capital gains are reinvested monthly. The *Monthly Gross Return USD* is based on the *Monthly Return USD* and adds the most recent net expense ratio. The *Net Assets - share class (Monthly) USD* is the monthly total net assets of a share class.

To distinguish between active and index funds, we make use of the *Index Fund* field. Those funds that track a particular index based on full replication or based on a representative sampling are flagged by Morningstar as index funds. Next, to filter the institutional and retail funds, we downloaded the field *Institutional*, which defines any fund as institutional if it either says "institutional" in the name of its share class, has a minimum investment above USD 100,000, or the prospectus says that it is for institutional investors only.

A.2 Data cleaning

For each fund, we retrieved its monthly net return, gross return, and total net assets, all in US dollars. We only included an observation if all three items were available. Often, and as reported in Berk and van Binsbergen (2015), we observe that net assets are reported quarterly or are missing for a specific month. In this case, we roll the assets under the assumptions of zero net flows, so as to increase the available data points and avoid disconnected time-series. Besides, for some institutional mutual funds, we observe zero fees because they are paid in separate contracts with the asset manager. Thus, we only include funds where the sum of the gross returns is larger than the sum of the net returns to exclude zero-fees funds. To avoid the incubation bias, we include funds only if they reach 5 million December 2016 US dollars in AUM.

We also see conversion errors, where funds assets suddenly increase by a high factor and then decrease again by a similar factor. First, we observe this behavior in emerging market currencies before 1999. Thus, we concentrate in the period before 1999 only on the developed currencies, Pound

Sterling, US Dollar, Euro, Singapore Dollar, Australian Dollar, Swedish Krona, South African Rand, Swiss Franc, Japanese Yen, New Zealand Dollar, Canadian Dollar, Norwegian Krone, Danish Krone. Also, we see that for some funds, the assets change by a factor higher than 100 and decrease in the next period to the same level as before the outlier. For these cases, we smooth the net assets over time if we see that the assets change by a factor higher than 10 and we decrease them in the next two periods by a factor of more than 0.5. But there are funds where this increase is verified by attaining the same fund levels in the future. Therefore, we only correct the assets if the same level is not exceeded in its future assets.

We also delete obvious mistakes, such as when an index fund shows high fees in the past and suddenly changes to a low fee. In this case, we keep only the low fee period, since we interpret this as being that either the fees were not correct or the fund changed from active to index. When we build the value-weighted portfolio for the investment categories, we also remove funds that show a beta below 0.05 relative to the average return of all the funds within the same investment category. Because of the low sensitivity to the average fund, these funds are not following a strategy similar to that of the rest of the group.

A.3 Aggregation of the share classes

Each line in the Morningstar dataset corresponds to a share class. In all, we obtain 435,453 lines of different share classes. Thus, we must aggregate the same share classes to avoid multiple tries of investment strategies by the same provider. First, we tried to use the fields *Administrator* and *Ticker of Fund's Oldest Share Class*; however, they are often missing. For this reason, we aggregated alphabetically subsequent mutual funds that are in the same Morningstar Category with the corresponding *Index Fund* flag and have a similar name. While Berk and van Binsbergen (2015) use the last word of the fund's name for the share class, we use the ratio provided by the SequenceMatcher of the difflib library in Python, which is based on the algorithm developed by Ratcliff and Metzener (1988) and, additionally, cleans the "junk" elements. We define two names to be similar if this ratio is above 0.8.

A.4 Summary statistics of investment categories

Table 1 shows the summary statistics of the cross-sectional monthly attributes across asset classes. Table A.1 provides a more detailed view of all the investment categories, where we find both index and active mutual funds. For the active funds, we analyze a total of 14,969 institutional and 46,300 retail funds, of which 56,136 are active funds and 5,133 are index funds. In general, there are fewer index funds, but they have higher average total net assets (TNA) and net returns, and also lower fees and about the same average number of years in the database. As expected, the institutional funds charge lower fees than their retail counterparts.

[Table A.1 about here.]

B The robust alpha test

Consider a fund with time-t return y_t and a set of K benchmark factor returns x_{tk}, $k = 1, ..., K$. A total of T returns are observed. We assume that these observations are generated by a stationary multivariate return distribution with mean vector μ and covariance matrix Σ:

$$\mu = \begin{pmatrix} \mu_y \\ \mu_{x_1} \\ \vdots \\ \mu_{x_K} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx_1} & \cdots & \sigma_{yx_K} \\ \sigma_{x_1y} & \sigma_{x_1}^2 & \cdots & \sigma_{x_1x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_Ky} & \sigma_{x_Kx_1}^2 & \cdots & \sigma_{x_K}^2 \end{pmatrix}, \tag{B.1}$$

with the observed means $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$. By defining a vector $\mu_X = (0, E[x_1], \cdots, E[x_K])'$, we can express the fund's alpha as

$$\alpha = E[y] - \mu_X' \Sigma_{XX}^{-1} y_X, \tag{B.2}$$

with

$$\Sigma_{XX} = \begin{pmatrix} 1 & E[x_1] & E[x_2] & \cdots & E[x_K] \\ E[x_1] & E[x_1^2] & E[x_1 x_2] & \cdots & E[x_1 x_K] \\ \vdots & \vdots & & \ddots & \vdots \\ E[x_K] & E[x_K x_1] & E[x_K x_2] & \cdots & E[x_K^2] \end{pmatrix} \quad \text{and} \quad y_X = \begin{pmatrix} E[y] \\ E[x_1 y] \\ \vdots \\ E[x_K y] \end{pmatrix}. \tag{B.3}$$

Then, we test for the hypothesis

$$H_0: \quad \alpha = 0 \qquad \text{against} \qquad H_1: \quad \alpha \neq 0. \tag{B.4}$$

Furthermore, we define $\zeta_k = E[y x_k]$, $\gamma_k = E[x_k^2]$, $\xi_{kj} = E[x_k x_j]$, $j > k$, and the combined vector $\nu = (\mu_y, \cdots, \mu_{x_k}, \cdots, \zeta_k, \cdots, \gamma_k, \cdots, \xi_{kj}, \cdots)' \in \mathbb{R}^{1+3k+k(k-1)/2}$ with sample counterpart $\hat{\nu}$. Now, we can express the true alpha as a function f of ν:

$$\alpha = E[y] - \mu_X' \Sigma_{XX}^{-1} y_X = f(\nu); \tag{B.5}$$

and the estimated alpha as function of $\hat{\nu}$: $\hat{\alpha} = f(\hat{\nu})$. As mentioned in Ledoit and Wolf (2008), under mild regularity conditions,

$$\sqrt{T}(\hat{\nu} - \nu) \xrightarrow{d} N(0, \Psi), \tag{B.6}$$

where Ψ is an unknown symmetric positive semi-definite matrix. By the delta method, we obtain

$$\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \nabla' f(\nu) \Psi \nabla f(\nu)) \tag{B.7}$$

with

$$\nabla' f(\nu) = \left(\frac{\partial f(v)}{\partial \mu_y}, \cdots, \frac{\partial f(v)}{\partial \mu_{x_k}}, \cdots, \frac{\partial f(v)}{\partial \zeta_k}, \cdots, \frac{\partial f(v)}{\partial \gamma_k}, \cdots, \frac{\partial f(v)}{\partial \xi_{kj}}, \cdots \right)'. \tag{B.8}$$

Given a consistent estimator $\hat{\Psi}$ of Ψ, we can compute a standard error for $\hat{\alpha}$ by

$$s(\hat{\alpha}) = \sqrt{\frac{\nabla' f(\nu) \hat{\Psi} \nabla f(\nu)}{T}}. \tag{B.9}$$

To test the null hypothesis in Equation (B.4), we focus on the bootstrap inference for time-series data outlined in Ledoit and Wolf (2008). In particular, we denote the optimal block length by b and define $l = \text{floor}(T/b)$. As shown in Kuensch and Goetze (1996), the bootstrapped estimator of $\hat{\Psi}^*$ is

$$\hat{\Psi}^* = \frac{1}{l}\sum_{j=1}^{l} \eta_j \eta_j', \tag{B.10}$$

where

$$\begin{aligned} z_t^* &= \left(y_t^* - \hat{\mu_y}^*, ..., x_{tk}^* - \hat{\mu_x}^*, ..., y_t x_{tk} - \hat{\zeta}_k^*, ..., {x_{tk}^*}^2 - \hat{\gamma}_k^*, ..., x_{tk}^* x_{tj}^* - \hat{\xi}_{kj}^*, ...\right), \\ \eta_j &= \frac{1}{\sqrt{b}}\sum_{t=1}^{b} z_{(j-1)b+t}. \end{aligned} \tag{B.11}$$

Next, the studentized statistics are

$$\tilde{d}_m^* = \frac{|\hat{\alpha}_m^* - \hat{\alpha}|}{s\left(\hat{\alpha}_m^*\right)}, \tag{B.12}$$

and the p-value is

$$PV = \frac{\{\tilde{d}_m^* \geq \hat{d}\} + 1}{M + 1}, \tag{B.13}$$

where $\hat{d}$ is the original studentized test statistic that was computed from the observed returns. We use Newey–West standard errors to calculate the original standard errors. Regarding the optimal block, we suggest using either the method of Politis and White (2004) and Patton et al. (2009) for the univariate case, or the method of Ledoit and Wolf (2008) for the bivariate case. For our empirical analysis, we would like to compare up to 30 different investment categories and, so far, there is no available method to make this comparison. Consequently, we will further discuss the optimal block size to use in our simulations in Appendix C.

C Accuracy of the robust alpha test

We now present the results of a simulation study to show the difference between our robust alpha test and the standard hypothesis tests.[42] For this purpose, we first simulate a single hypothesis setting.

[42] Other papers that extensively use bootstrap techniques often do not perform such a simulation study to validate their approach.

For realistic time-series, we select the first ten US mutual funds of the Morningstar database within the category "US Equity Large Cap Blend" that offer the entire return history from 1992 to 2016 (T=300). As benchmark models, we focus on the one-factor "CAPM," i.e., the market excess return, the three-factor "FF3," and the five-factor "FF5" model. For the data generating process (DGP), we sample from the realized returns with a circular block bootstrap and block sizes of 1, 3, and 6. We selected this grid of block sizes based on our analysis in Section 2, where we observe that most of the optimal circular block sizes range from one to six. This grid corresponds to time periods of one, three, and six months. The block sizes of three and six are the ones that take the evidence of serial dependence from Section 2 into account. A block size of one generates independent data, and we employ this block size only for reasons of comparison. For each fund, we simulate 1,000 paths and set the alpha under the null hypothesis to the true observed alpha of the data. The bootstrapped p-values (Boot) are then calculated as illustrated in Appendix B by employing $M = 1{,}000$ and the optimal block size by the method of Politis and White (2004) and the correction of Patton et al. (2009). We compare the robust p-values with those from the standard inference methods; that is, based on the normal distribution (Standard), Newey–West (NW), and HC3 standard errors.

[Table C.1 about here.]

Table C.1 shows the empirical rejection probabilities of the falsely rejected null hypothesis compared to the nominal levels $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$. Because the null hypothesis is true for all the simulations, the true rejection probabilities should be equal to the nominal levels of the test. If a test shows a higher percentage of rejections, then we regard this test as too liberal. While we observe that the standard inference tests based on the normal distribution, the Newey–West, and HC3 standard errors are too liberal in rejecting the null hypothesis, the bootstrapped solution (Boot) presented in the previous section is close to the nominal levels. We highlight in bold the empirical rejection probabilities that are closest to the desired level. We observe the HC3 standard errors to be in some cases closer to the desired level than are those of the block bootstrapped method, but only in the case where we apply the standard but less realistic bootstrap with a block size of one where we lose any dependence over time. However, as we demonstrate in Section 2, the optimal block size, and thus a realistic assumption for the DGP is, in general, around three or six, for which our

bootstrapped test is tailored to be more accurate.[43]

Since there is still the open question of the optimal block size in the multiple hypothesis setting when controlling the FWER, as illustrated in Romano and Wolf (2005a,b, 2016), we conduct a second simulations study. Unlike the single mutual fund analysis, where we regard each fund in isolation and then apply the multiple hypothesis framework of Barras et al. (2010), in this case, we must consider the cross-dependence structure, and jointly sample the funds and benchmark returns. For this purpose, we focus on the 17 portfolios within the "Inv. Categories" setting from Section 2 with the investable one-factor benchmark model that is based on the value-weighted return of index funds. Also, instead of calculating the Type I Errors as in the single hypothesis setting, we compute the empirical rejection probabilities based on the FWER, as illustrated in Romano and Wolf (2005a,b, 2016). To find the optimal block size that is closest to the nominal levels of the test, we focus on the following grid of block sizes: 1, 3, 6, 9, and 12. Regarding the DGP, we keep the grid from our first simulation study.

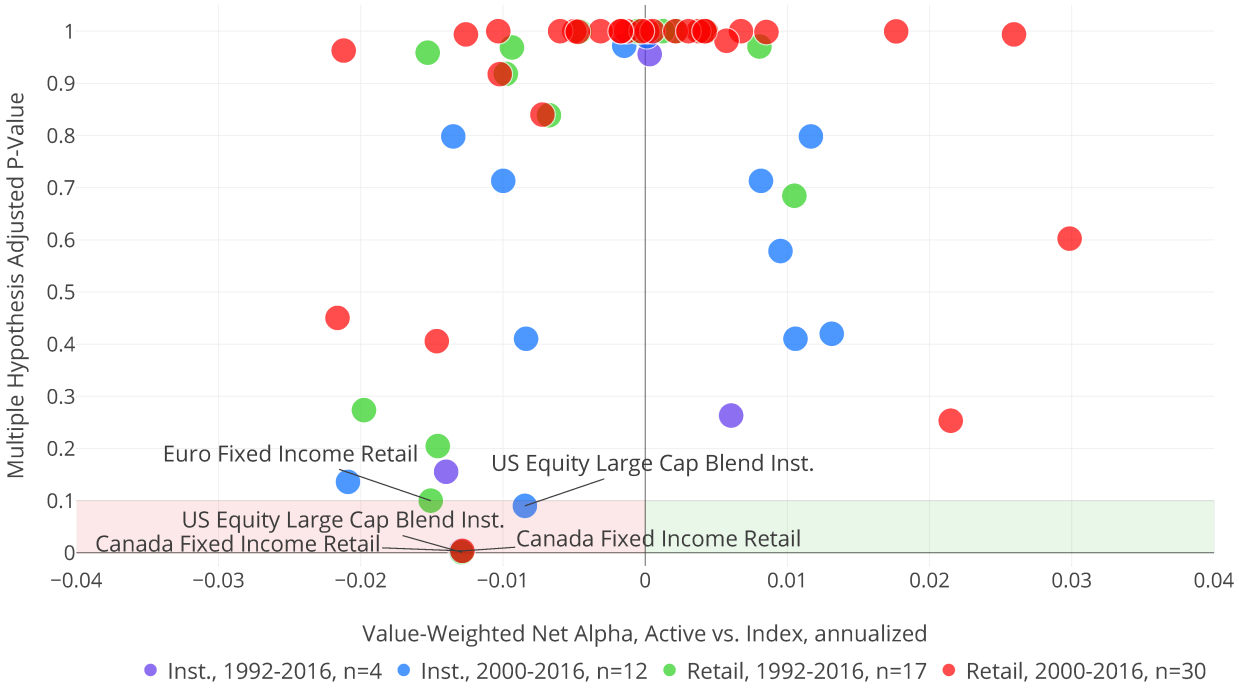
[Table C.2 about here.]

Table C.2 shows the empirical rejection probabilities based on the FWER. Likewise, for the FWER, we find the bootstrapped robust alpha test to achieve the desired levels at optimal block sizes three or six. Given that for a block size of three we observe accurate rejection probabilities, in the multiple comparisons of portfolios, we will, in the remainder of the paper, present the results based on the optimal block size of three. Finally, the more conservative block sizes six and nine are applied for robustness checks.

[43] A similar observation was also made in Ledoit and Wolf (2008) for testing the Sharpe Ratio and in Ledoit and Wolf (2011) for the variance.

Figure 1: Value-weighted alpha of active mutual funds within investment categories

Multiple hypotheses adjusted p-value (y-axis) and annualized value-weighted alpha of active versus index funds (x-axis) for all investment categories as defined by the "Global Category" of Morningstar. Top (bottom): analysis after (before) management fees. We form the four groups with the combinations retail and institutional as well as the periods 1992–2016 and 2000–2016.



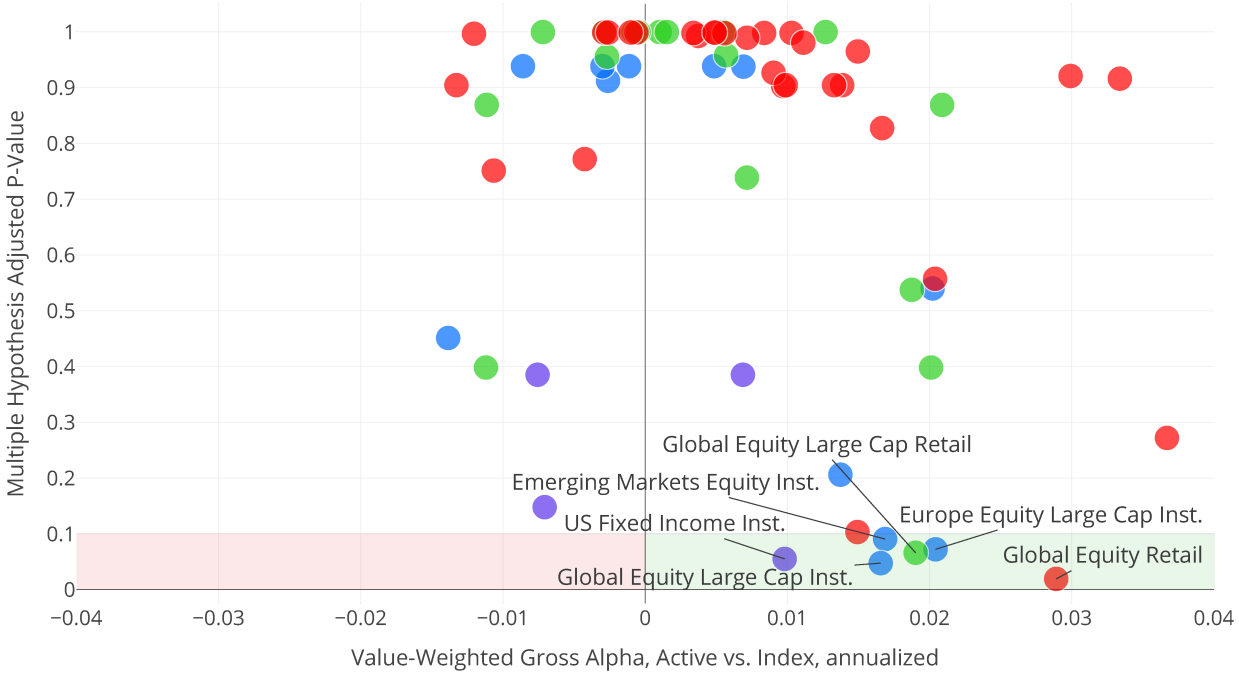


Figure 2: Aggregated value-weighted alpha of active minus index

Cumulated logarithmic alphas for the active equity (top) and fixed income (bottom) mutual funds. The alpha is the value-weighted return of the active funds against the value-weighted return of the index funds within the same investment category. The figure shows the aggregated alpha with equal-weights (EW) and value-weights (VW) across the Morningstar investment categories. We analyze both institutional and retail funds. We also regard the portfolios before (Gross) and after (Net) costs. We include all mutual funds within Morningstar where net and gross returns and assets under management are available, and where we have at least one index fund within the same investment category. The analysis is in US dollars.

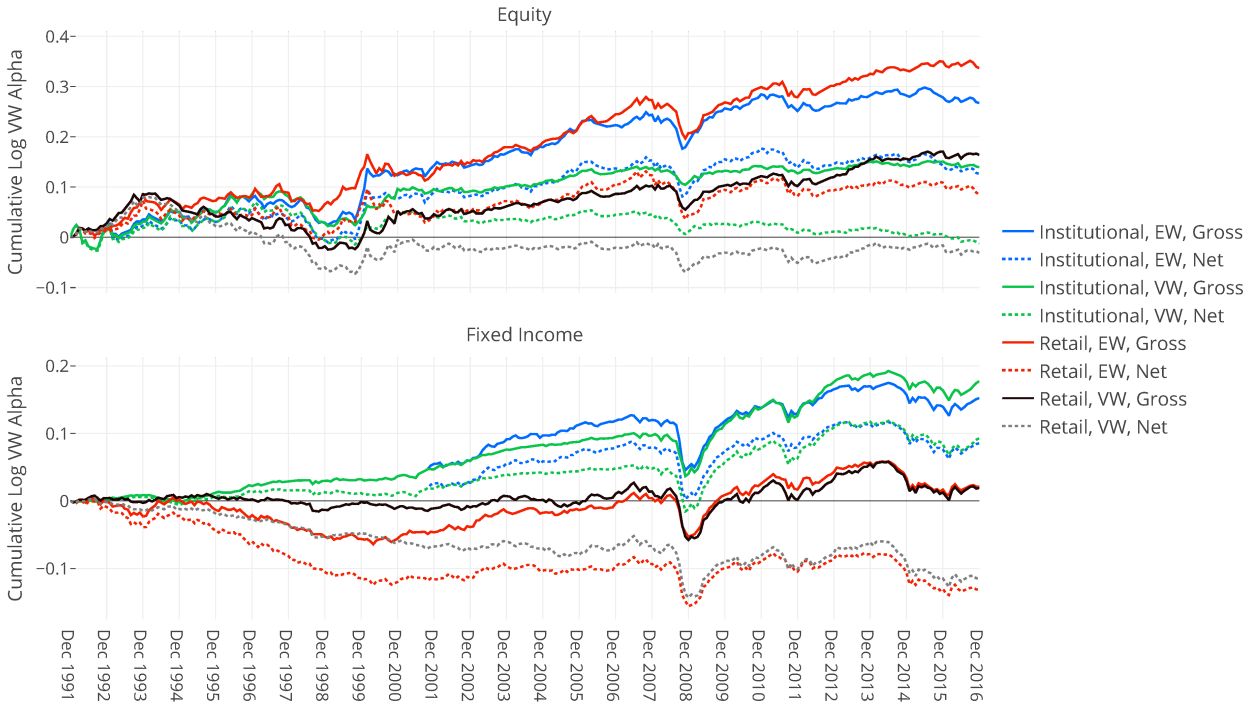


Figure 3: Active equity fees of young and old funds in the US

Average active fee over the last year (top) of US equity funds with a track record of more than five years (>5y) and with a track record of at most one year (<1y), and the percentage of index funds (bottom) within all US equity mutual funds. We distinguish between retail and institutional funds.

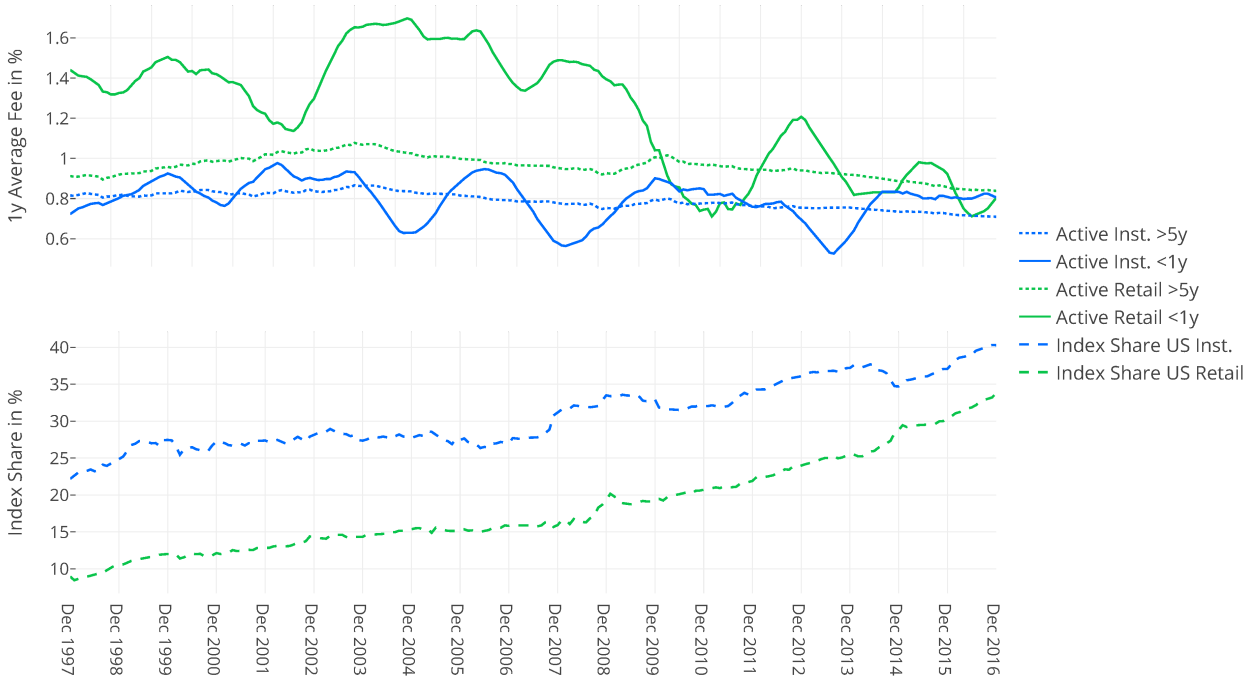


Table 1: Mutual fund database summary statistics

Total number, average number (Avg Number), the average total net assets in million USD (Avg TNA), average annual net return in USD (Avg Net Ret), average annual fee in USD (Avg Fees ann), and the average years of a fund in the database (Avg Years) over the time period from December 1991 to December 2016 of all available funds in the Morningstar database flagged by Open-End or Exchange-Traded funds. We only include funds within the "Global Broad Category Group" equity (Equity) and fixed income (Fixed Income) for which we provide the category statistics. The average corresponds to the mean of cross-sectional monthly attributes.

in USD		Total Number		Avg Number		Avg TNA		Avg Net Ret		Avg Fees ann		Avg Years	
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*
Equity	*Inst.*	8,488	691	2,506.1	199.5	255.3	910.0	8.85%	8.95%	0.86%	0.15%	7.4	7.2
	Retail	26,741	3,551	9,147.8	950.6	453.6	732.2	8.16%	8.27%	1.18%	0.31%	8.6	6.7
Fixed	*Inst.*	5,566	224	1,440.3	57.8	333.2	663.0	4.97%	5.19%	0.54%	0.19%	6.5	6.5
Income	*Retail*	15,341	667	4,545.6	152.6	346.7	817.6	4.87%	5.20%	0.88%	0.25%	7.4	5.7

Table 2: Dependence analysis

Results of dependence analysis. Panel A shows the total number of funds within each category and the percentage of mutual funds with a significant serial dependence according to the Ljung–Box (LJ) and Genest and Rémillard (2004) (GR) tests. For the latter, we use 1,000 simulations. Panel B shows the average correlations (AvgCorr) and the p-values for the cross-sectional dependence test of Pesaran (2004). We analyze the residuals of single mutual funds (Single) and portfolios of mutual funds (Portfolio), split into equity (Equity) and fixed income (FixedInc) mutual funds. For the portfolio of mutual funds, we additionally analyze all 63 investment category portfolios (InvCat) from Figure 1. For the benchmark model, we use the investable one-factor model with the value-weighted return of the index funds within the same category as the analyzed single mutual fund or portfolio of mutual funds (Inv), the equity five-factor model (FF5) with the regional factors "market," "size," and "value" of Fama and French (1992), and also "momentum" of Jegadeesh and Titman (1993) and "betting against beta" of Frazzini and Pedersen (2014), and the regional four-factor fixed income model (FI4) with the "shift," "twist" and "butterfly" factors, as well as the difference between the BBB and AAA credit spread.

Panel A: Serial dependence

		Single				Portfolio		
		Equity		FixedInc		InvCat	Equity	FixedInc
		Inv	FF5	Inv	FI4	Inv	FF5	FI4
Retail	Total	24,456	12,816	14,579	4,719	47	5	4
	GR<5%	17%	20%	22%	16%	28%	60%	25%
	LJ<5%	19%	23%	19%	18%	19%	80%	25%
Inst.	Total	7,025	3,817	4,815	1,528	16	5	4
	GR<5%	14%	15%	15%	18%	31%	0%	0%
	LJ<5%	15%	18%	19%	21%	44%	20%	25%

Panel B: Cross-sectional dependence

		Single				Portfolio		
		Equity		FixedInc		InvCat	Equity	FixedInc
		Inv	FF5	Inv	FI4	Inv	FF5	FI4
Retail	AvgCorr	0.04	0.22	0.11	0.62	0.14	0.11	0.44
	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Inst.	AvgCorr	0.06	0.21	0.15	0.60	0.12	0.06	0.41
	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Overview of the statistical methodology

Whether different statistical methodologies can cope with potential dependencies in the data, serial dependence (Time) and cross-sectional dependence (Cross). To obtain the p-values needed for applying the two multiple-hypothesis tests, the FDR of Barras et al. (2010) and the FWER of Romano and Wolf (2016), use either the normal distribution (Standard), Newey–West (NW), HC3, the standard resampling (Robust-SR), or the block resampling (Robust-BR) standard errors. For the NW standard errors we use the automatic bandwidth selection procedure described in Newey and West (1994) based on the Bartlett kernel. The HC3 standard errors are consistent even in the presence of heteroscedasticity of an unknown form. The last two columns show what the data tells us concerning the dependencies for single funds and for the fund portfolios.

	Dependence:				Dependence:	
	Serial	Cross-sectional			Serial	Cross-sectional
Methodology:		FDR	FWER	Evidence/Data:		
Standard:	×	✓	not applicable	Single funds:	✓	✓
NW:	✓	✓	not applicable	Portfolios:	✓	✓
HC3:	×	✓	not applicable			
Standard-RS:	×	✓	✓			
Block-RS:	✓	✓	✓			

Table 4: Single mutual equity and fixed income funds against multi-factor benchmark

Proportion (in percentages) of zero, positive, and negative alpha funds, and significant p-values based on the FDR at the 10% significance level for funds with a positive and negative alpha. The alpha is after fees and relative to a multi-factor benchmark and across the equity investment regions US, Global, Europe, Japan, and Asia ex-Japan, as well as for the USD, CHF, EUR, and GBP fixed income markets. The results are based on the method of Barras et al. (2010), while we apply our robust alpha test to compute the single mutual funds' p-values. We show the results for the five-factor equity benchmark model based on the regional MKT, SMB, HML, WML, and BAB factors. For the fixed income benchmark, we include the four local factors "shift," "twist," and "butterfly," as well as the AAA–BBB credit spread.

Panel A: Retail funds

		Equity - 5 factors						Fixed Income - 4 factors				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	55.1	39.6	66.2	67.9	83.0	62.3	38.9	71.0	77.8	83.3	67.8
	Positive alpha	0.0	0.0	3.0	5.7	0.0	1.7	23.3	3.3	22.2	16.7	16.4
	Negative alpha	44.9	60.4	30.8	26.4	17.0	35.9	37.8	25.7	0.0	0.0	15.9
	FDR 10 alpha>0	0.0	0.1	0.0	0.0	0.0	0.0	18.3	0.0	0.0	0.0	4.6
	FDR 10 alpha<0	30.9	64.9	16.5	15.2	0.0	25.5	36.9	0.9	0.0	0.0	9.4
Index	*Zero alpha*	61.9	30.1	76.5	73.7	100.0	68.4	41.6	93.3	71.6	100.0	76.6
	Positive alpha	0.0	0.0	3.6	5.9	0.0	1.9	29.2	6.7	28.4	0.0	16.1
	Negative alpha	38.1	69.9	19.9	20.4	0.0	29.7	29.2	0.0	0.0	0.0	7.3
	FDR 10 alpha>0	0.0	0.0	0.0	1.3	0.0	0.3	5.0	0.0	0.0	0.0	1.3
	FDR 10 alpha<0	25.3	79.6	0.0	6.6	0.0	22.3	15.0	0.0	0.0	0.0	3.8

Panel B: Institutional funds

		Equity - 5 factors						Fixed Income - 4 factors				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	69.3	53.5	78.5	88.2	97.4	77.4	38.5	77.4	60.1	82.7	64.7
	Positive alpha	0.0	0.0	8.2	9.4	0.0	3.5	40.7	22.6	39.9	17.3	30.1
	Negative alpha	30.7	46.5	13.3	2.4	2.6	19.1	20.8	0.0	0.0	0.0	5.2
	FDR 10 alpha>0	0.0	0.0	0.0	0.0	0.0	0.0	42.1	0.0	0.2	0.0	10.6
	FDR 10 alpha<0	3.0	38.0	0.6	0.5	0.0	8.4	18.1	0.0	0.0	0.0	4.5
Index	*Zero alpha*	66.9	55.7	91.9	92.5	90.6	79.5	57.5	71.4	56.5	95.0	70.1
	Positive alpha	0.0	0.0	6.8	0.0	0.0	1.4	21.3	26.2	43.5	0.0	22.7
	Negative alpha	33.1	44.3	1.4	7.5	9.4	19.1	21.3	2.4	0.0	5.0	7.2
	FDR 10 alpha>0	0.0	0.0	0.0	5.0	0.0	1.0	5.0	0.0	0.0	0.0	1.3
	FDR 10 alpha<0	17.2	37.1	0.0	0.0	0.0	10.9	15.0	0.0	0.0	0.0	3.8

Table 5: Single mutual equity and fixed income funds against investable benchmark

Proportion (in percentage numbers) of zero, positive, negative alpha funds, and significant p-values based on the FDR at the 10% significance level for funds with a positive and negative alpha. The alpha is after fees and relative to the investable value-weighted portfolio of index funds within the same Morningstar investment category. The results are based on the method of Barras et al. (2010), while single mutual funds' p-values are derived from our robust alpha test.

Panel A: Retail funds

		Equity - Investable						Fixed Income - Investable				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	71.9	78.3	78.0	87.3	94.3	82.0	63.3	58.2	53.1	95.9	70.5
	Positive alpha	0.0	12.2	8.6	0.0	0.0	4.2	19.1	0.0	0.0	0.0	4.7
	Negative alpha	28.1	9.5	13.4	12.7	5.7	13.9	17.6	41.8	46.9	4.1	24.9
	FDR 10 alpha>0	0.0	1.4	0.0	0.0	0.4	0.3	6.7	0.0	0.0	0.0	1.4
	FDR 10 alpha<0	6.3	0.8	2.1	2.0	0.0	2.2	8.3	7.4	35.3	0.0	10.6
Index	*Zero alpha*	64.2	62.8	88.0	100.0	100.0	83.0	58.2	89.0	73.8	97.1	80.2
	Positive alpha	2.5	30.3	4.1	0.0	0.0	7.4	21.2	0.0	11.1	0.0	7.9
	Negative alpha	33.3	6.9	7.9	0.0	0.0	9.6	20.7	11.0	15.1	2.9	11.8
	FDR 10 alpha>0	0.0	22.0	0.0	0.0	0.0	4.4	1.5	0.0	0.0	0.0	1.2
	FDR 10 alpha<0	26.1	1.7	0.0	0.7	0.0	5.7	2.0	2.4	0.0	0.0	2.0

Panel B: Institutional funds

		Equity - Investable						Fixed Income - Investable				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	80.1	81.2	83.5	89.6	100.0	86.9	55.2	75.4	60.1	100.0	75.5
	Positive alpha	0.0	14.3	13.3	2.7	0.0	6.0	33.5	0.0	0.0	0.0	7.9
	Negative alpha	19.9	4.5	3.2	7.7	0.0	7.1	11.4	24.6	39.9	0.0	16.6
	FDR 10 alpha>0	0.0	1.2	0.0	0.0	0.0	0.2	18.2	0.0	0.0	0.0	3.7
	FDR 10 alpha<0	0.2	0.3	0.0	0.0	0.4	0.2	4.9	0.0	13.2	0.0	3.7
Index	*Zero alpha*	53.1	44.6	92.9	100.0	94.4	77.0	95.6	90.8	78.6	100.0	88.4
	Positive alpha	6.1	44.1	7.1	0.0	2.8	12.0	2.2	0.0	0.0	0.0	2.8
	Negative alpha	40.8	11.4	0.0	0.0	2.8	11.0	2.2	9.2	21.4	0.0	8.8
	FDR 10 alpha>0	0.0	48.5	0.0	0.0	0.0	9.7	2.2	0.0	0.0	0.0	2.4
	FDR 10 alpha<0	34.6	2.0	0.0	0.0	0.0	7.3	0.0	0.0	0.0	0.0	1.5

Table 6: Performance drivers of active minus index

Results from regressing the difference between value-weighted active investing and index investing, before fees. For the benchmark model, we include the difference between the VIX index and the regional equity model with the regional MKT, SMB, HML, WML, and BAB factors. For the regional fixed income model, we add the VIX index to the four local factors shift, twist, and butterfly (BFLY), as well as the AAA–BBB credit spread (SPR). Coefficient estimates are multiplied by 100 and HC3 standard errors are in parentheses. By *, **, and *** we denote p-values below 0.1, 0.05, and 0.01, respectively. The last rows report the adjusted R^2 values.

Panel A: Equity funds

			Retail		*Asia ex-*			Institutional		*Asia ex-*
	US	*Global*	*Europe*	*Japan*	*Japan*	*US*	*Global*	*Europe*	*Japan*	*Japan*
Const.	−0.01	0.27***	0.04	0.10	−0.15	−0.06*	0.15***	0.07	0.07	0.17
	(0.04)	(0.06)	(0.05)	(0.09)	(0.20)	(0.03)	(0.05)	(0.07)	(0.07)	(0.15)
MKT	−2.41**	−7.07***	4.04**	1.88	3.97	−0.25	−2.80**	1.34	2.37	−4.71*
	(1.09)	(1.90)	(1.60)	(2.56)	(6.05)	(1.08)	(1.28)	(1.90)	(1.62)	(2.49)
SMB	13.77***	14.87***	23.14***	7.58**	6.73	26.39***	14.72***	35.59***	15.63***	22.50***
	(1.84)	(2.93)	(3.10)	(3.77)	(4.94)	(1.53)	(2.04)	(4.33)	(3.29)	(4.98)
HML	−6.93***	−3.85	−8.32***	−12.41***	13.73***	−3.95**	−0.57	2.15	−1.19	22.73***
	(1.74)	(2.69)	(2.29)	(3.27)	(4.95)	(1.77)	(2.25)	(3.09)	(3.07)	(5.42)
WML	2.73*	4.78***	2.75*	12.47***	−0.47	2.09*	2.01	5.88***	−1.54	0.20
	(1.42)	(1.61)	(1.57)	(3.26)	(4.25)	(1.26)	(1.28)	(1.83)	(2.46)	(3.79)
BAB	0.41	−9.50***	−6.51***	−3.00	5.51	0.39	0.28	−6.59***	−8.60***	−12.41**
	(1.52)	(2.61)	(1.75)	(2.55)	(6.61)	(1.34)	(1.99)	(2.28)	(2.39)	(5.84)
Δ *VIX*	−3.46***	−7.16***	−2.94*	−3.38	7.63	−2.39**	−3.57***	−3.68**	−3.48*	−4.02
	(0.98)	(2.00)	(1.75)	(2.36)	(6.11)	(1.05)	(1.34)	(1.69)	(1.80)	(3.51)
R^2	0.51	0.20	0.26	0.25	0.08	0.73	0.18	0.45	0.22	0.21

Panel B: Fixed income funds

	Retail				Institutional			
	USD	*CHF*	*EUR*	*GBP*	*USD*	*CHF*	*EUR*	*GBP*
Const.	0.05***	0.02	0.03*	0.13**	0.06***	−0.00	0.02	0.08*
	(0.01)	(0.02)	(0.02)	(0.05)	(0.02)	(0.01)	(0.03)	(0.04)
SHIFT	−0.91***	−2.21***	−2.73***	−2.48***	−0.71***	−0.50***	−2.26***	−0.23
	(0.09)	(0.27)	(0.17)	(0.48)	(0.17)	(0.09)	(0.24)	(0.36)
TWIST	−0.33*	−0.82	−0.09	−1.40	−1.33***	−0.37**	−0.27	1.11*
	(0.17)	(0.57)	(0.26)	(1.45)	(0.29)	(0.16)	(0.45)	(0.59)
BFLY	−0.30	0.99*	0.76	2.40	−0.72	−0.31	0.85	1.09
	(0.32)	(0.51)	(0.60)	(2.26)	(0.59)	(0.28)	(0.96)	(1.23)
SPR	1.27***	0.30**	0.27	3.81***	1.13***	0.12**	−0.11	1.60***
	(0.21)	(0.14)	(0.22)	(0.90)	(0.33)	(0.06)	(0.30)	(0.47)
Δ *VIX*	−1.86**	−2.97***	−1.17	−5.34*	−1.40	−0.17	−2.86**	−0.90
	(0.73)	(1.01)	(0.81)	(2.91)	(1.19)	(0.24)	(1.20)	(1.78)
R^2	0.82	0.59	0.81	0.72	0.63	0.25	0.63	0.49

Table 7: One-year persistence of the alpha after fees

Annualized alpha after fees (in %), the corresponding block-bootstrapped multiple hypothesis adjusted p-value (in brackets), and the beta for the value-weighted performance of active mutual funds benchmarked against the value-weighted performance of index funds. In Panel A, each row from 100% to 10% corresponds to the value-weighted portfolio including only the x% best active mutual funds of the past year based on the t-value for the alpha. In Panel B, we report the same numbers but for the x% worst active mutual funds of the past year. Every December, we rebalance the momentum portfolios. The data sample ranges from 1993 to 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

Panel A: Value-weighted performance of the x% best performing funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.12	−0.12	−0.10	−0.01	0.07	0.07	0.11	0.40	0.62
			(0.78)	(0.94)	(0.94)	(0.95)	(0.98)	(0.97)	(0.97)	(0.97)	(0.73)	(0.65)
		beta	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98
	Retail	*alpha*	−0.60	−0.46	−0.44	−0.41	−0.44	−0.40	−0.31	−0.27	−0.13	−0.02
			(0.30)	(0.52)	(0.52)	(0.53)	(0.52)	(0.53)	(0.64)	(0.71)	(0.91)	(0.97)
		beta	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Fixed Income	Inst.	*alpha*	0.26	0.32	0.35	0.33	0.30	0.28	0.27	0.22	0.13	0.16
			(0.75)	(0.68)	(0.59)	(0.59)	(0.66)	(0.68)	(0.68)	(0.75)	(0.75)	(0.75)
		beta	0.88	0.87	0.86	0.86	0.85	0.85	0.85	0.84	0.80	0.78
	Retail	*alpha*	−0.75	−0.73	−0.70	−0.70	−0.67	−0.62	−0.56	−0.55	−0.58	−0.40
			(0.21)	(0.24)	(0.26)	(0.26)	(0.25)	(0.26)	(0.26)	(0.26)	(0.22)	(0.26)
		beta	0.97	0.96	0.96	0.95	0.94	0.93	0.92	0.92	0.91	0.89

Panel B: Value-weighted performance of the x% worst performing funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.24	−0.30	−0.32	−0.40	−0.51	−0.51	−0.65	−1.01	−0.94
			(0.64)	(0.64)	(0.64)	(0.64)	(0.60)	(0.56)	(0.58)	(0.46)	(0.24)	(0.43)
		beta	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	Retail	*alpha*	−0.60	−0.63	−0.72	−0.77	−0.80	−0.79	−0.88	−1.04	−1.08	**−1.39**
			(0.27)	(0.27)	(0.26)	(0.27)	(0.27)	(0.27)	(0.27)	(0.19)	(0.17)	(0.04)
		beta	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.98
Fixed Income	Inst.	*alpha*	0.26	0.26	0.13	0.23	0.14	0.13	0.16	0.14	−0.08	−0.26
			(0.77)	(0.78)	(0.96)	(0.85)	(0.96)	(0.96)	(0.94)	(0.96)	(0.96)	(0.80)
		beta	0.88	0.89	0.90	0.91	0.91	0.93	0.94	0.96	0.96	0.98
	Retail	*alpha*	−0.75	−0.77	−0.84	−0.93	−0.96	−0.96	−0.96	−0.84	−0.93	*−0.84*
			(0.17)	(0.17)	(0.17)	(0.17)	(0.15)	(0.15)	(0.14)	(0.17)	(0.11)	(0.09)
		beta	0.97	0.98	0.99	1.01	1.02	1.02	1.02	1.03	1.02	1.00

Table 8: Bivariate sorts on the performance and size of the past year

Annualized alpha after fees (in %) and block-bootstrapped multiple hypothesis adjusted p-value (in brackets) for the value-weighted performance of active mutual funds against the investable benchmark. The portfolios of active mutual funds are double-sorted based on the performance (rows) and the size (columns) of the past year. The nine portfolios arise from the 30th and 70th percentiles within the investment category of each fund. For each panel we distinguish between equity (top) and fixed income (bottom) funds. We rebalance the portfolios every year starting in December 1992 to December 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

		Institutional				Retail		
		Small	Medium	Big		Small	Medium	Big
Equity	Winner	0.71	0.83	−0.05	Winner	0.09	−0.45	−0.28
		(0.76)	(0.54)	(0.96)		(0.89)	(0.60)	(0.74)
	Average	0.65	−0.11	−0.22	Average	−1.15	−0.96	−0.66
		(0.74)	(0.96)	(0.93)		(0.17)	(0.17)	(0.44)
	Loser	−0.76	−0.42	−0.65	Loser	*−1.78*	*−1.53*	−0.99
		(0.69)	(0.93)	(0.70)		(0.07)	(0.05)	(0.30)
Fixed Income	Winner	0.59	−0.25	0.25	Winner	−0.61	−0.49	−0.56
		(0.60)	(0.86)	(0.84)		(0.20)	(0.27)	(0.27)
	Average	−0.31	−0.14	0.37	Average	−1.00	−0.86	−0.93
		(0.86)	(0.92)	(0.80)		(0.13)	(0.21)	(0.27)
	Loser	−0.66	−0.13	0.20	Loser	*−1.25*	−1.05	−0.79
		(0.63)	(0.92)	(0.92)		(0.06)	(0.12)	(0.27)

Table 9: Portfolios filtered by the fee of the past year.

Annualized alpha after fees (in %), the corresponding block-bootstrapped multiple hypothesis adjusted p-value (in brackets), and the beta for the value-weighted performance of active mutual funds against the investable benchmark. Panel A shows the value-weighted performance of portfolios consisting of the x% least expensive active mutual funds of the past year. Panel B shows the performance of the x% most expensive active mutual funds. Every December, the portfolios are rebalanced to exclude a certain percentage of active funds. The data sample ranges from 1993 to 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

Panel A: Value-weighted performance of the x% least expensive funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.13	−0.11	−0.05	−0.04	0.09	0.04	0.17	0.49	0.83
			(0.66)	(0.82)	(0.84)	(0.96)	(0.96)	(0.86)	(0.96)	(0.74)	(0.26)	(0.15)
		beta	1.00	1.00	0.99	1.00	0.99	0.99	1.00	1.00	0.99	0.97
	Retail	*alpha*	−0.60	−0.50	−0.44	−0.40	−0.33	−0.24	−0.15	−0.04	0.03	0.30
			(0.23)	(0.35)	(0.42)	(0.48)	(0.58)	(0.66)	(0.80)	(0.97)	(0.97)	(0.62)
		beta	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.97	0.96
Fixed Income	Inst.	*alpha*	0.26	0.32	0.33	0.33	0.35	0.33	0.35	0.39	0.25	0.01
			(0.69)	(0.61)	(0.61)	(0.56)	(0.52)	(0.52)	(0.46)	(0.37)	(0.61)	(0.96)
			0.88	0.88	0.87	0.87	0.86	0.86	0.86	0.86	0.81	0.72
	Retail	*alpha*	−0.75	−0.71	−0.66	−0.63	−0.61	−0.56	−0.55	−0.53	−0.42	−0.40
			(0.17)	(0.19)	(0.20)	(0.20)	(0.22)	(0.25)	(0.26)	(0.28)	(0.36)	(0.36)
			0.97	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.94	0.94

Panel B: Value-weighted performance of the x% most expensive funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.36	−0.52	−0.50	−0.48	−0.78	−0.65	−0.94	−0.88	−1.14
			(0.50)	(0.36)	(0.30)	(0.32)	(0.34)	(0.20)	(0.32)	(0.22)	(0.30)	(0.20)
		beta	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02	1.03	1.04
	Retail	*alpha*	−0.60	**−0.99**	**−1.28**	**−1.53**	**−1.70**	**−1.82**	**−1.94**	**−2.10**	**−2.48**	**−2.83**
			(0.15)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
		beta	0.99	1.00	1.01	1.01	1.01	1.02	1.02	1.02	1.02	1.01
Fixed Income	Inst.	*alpha*	0.26	0.31	0.22	0.10	0.10	0.15	0.06	0.08	−0.11	−0.26
			(0.66)	(0.60)	(0.77)	(0.95)	(0.95)	(0.92)	(0.97)	(0.97)	(0.97)	(0.84)
			0.88	0.89	0.89	0.90	0.90	0.92	0.94	0.94	0.94	0.88
	Retail	*alpha*	−0.75	−0.83	−0.91	−0.92	−0.95	−1.02	−1.02	−1.08	−1.37	−1.46
			(0.16)	(0.16)	(0.14)	(0.14)	(0.14)	(0.14)	(0.16)	(0.16)	(0.14)	(0.14)
			0.97	0.97	0.98	0.98	0.99	1.00	1.00	1.00	1.02	1.00

Table 10: Bivariate sorts on the performance and fee of the past year

Annualized alpha after fees (in %) and block-bootstrapped multiple hypothesis adjusted p-value (in brackets) for the value-weighted performance of active mutual funds against the investable benchmark. The portfolios of active mutual funds are double-sorted based on the performance (rows) and the fee (columns) of the past year. The nine portfolios arise from the 30th and 70th percentiles within the investment category of each fund. For each panel we distinguish between equity (top) and fixed income (bottom) funds. We rebalance the portfolios every year starting in December 1992 to December 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

		Institutional				Retail		
		High Fee	Medium	Low Fee		High Fee	Medium	Low Fee
Equity	Winner	−0.03	−0.19	0.44	Winner	*−1.78*	−1.03	0.14
		(0.99)	(0.99)	(0.92)		(0.08)	(0.16)	(0.93)
	Average	−1.15	−0.16	−0.14	Average	**−2.13**	*−1.24*	−0.18
		(0.33)	(0.99)	(0.99)		(0.00)	(0.06)	(0.93)
	Loser	−1.61	−1.41	0.25	Loser	**−2.38**	*−1.55*	−0.43
		(0.33)	(0.23)	(0.99)		(0.00)	(0.05)	(0.81)
Fixed Income	Winner	0.10	0.33	0.17	Winner	−0.85	−0.60	−0.42
		(1.00)	(0.77)	(0.91)		(0.37)	(0.35)	(0.37)
	Average	0.17	−0.01	0.49	Average	−1.20	−0.96	−0.74
		(1.00)	(1.00)	(0.57)		(0.35)	(0.22)	(0.37)
	Loser	−0.07	0.04	0.40	Loser	−1.17	−0.79	−0.71
		(1.00)	(1.00)	(0.91)		(0.22)	(0.35)	(0.37)

Table 11: Percentage of rejections under the standard and robust alpha tests

Rejections (in %) based on the normal distribution (Standard), Newey–West (NW), HC3, standard resampling (Robust-SR), and the block resampling (Robust-BR) standard errors. For the single mutual fund analysis, we show the percentage of rejections for all 52,526 active mutual funds for the single hypothesis and for the multiple hypothesis analysis based on controlling the FDR. For the portfolios of mutual funds, we compare the percentage of rejections for the 63 portfolios where we adjust for multiple tries based on the FWER. For the standard tests, the FWER method of Romano and Wolf (2016) is not applicable since it requires the bootstrapped t-values.

	Single Mutual Funds				**Portfolios of Mutual Funds**			
Single Hypothesis	Unskilled		Skilled		Unskilled		Skilled	
	<5%	<10%	<5%	<10%	<5%	<10%	<5%	<10%
Standard	11.2	16.2	5.4	8.1	17.5	20.3	7.8	11.1
NW	12.3	17.5	5.8	8.7	19.0	20.3	4.8	7.9
HC3	10.4	15.3	5.0	7.7	17.5	20.3	7.8	11.1
Robust - SR	11.1	16.4	5.3	8.1	19.0	20.3	4.7	7.9
Robust - BR	9.0	13.9	4.3	6.9	15.9	18.7	1.6	6.3
Multiple Hypothesis	FDR				FWER			
	Unskilled		Skilled		Unskilled		Skilled	
	<5%	<10%	<5%	<10%	<5%	<10%	<5%	<10%
Standard	4.1	7.9	1.4	2.2	not applicable			
NW	4.9	9.3	1.3	2.2	not applicable			
HC3	3.2	6.4	1.3	1.8	not applicable			
Robust - SR	3.3	7.0	0.9	1.6	6.4	9.5	0	0
Robust - BR	1.4	3.8	0.3	0.8	4.7	6.4	0	0

Table A.1: Summary statistics for mutual fund investment categories

Average number (Avg Number) of funds, average total net assets in million USD (Avg TNA), average annual net return in USD (Avg Net Ret), average number of years the fund is in the database (Avg Years), and the first appearance of an index fund, for the time period from December 1991 to December 2016, for all available investment categories (Global Category) within the Morningstar database. We only include funds within the "Global Broad Category Group" equity or fixed income that are flagged as "Open-End" or "Exchange-Traded" funds. The average corresponds to the mean of cross-sectional monthly attributes.

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
Mexico Fixed Income	*Retail*	182.3	2.0	233.7	137.4	-1.89%	1.28%	5.8	7.5	May 09
	Inst.	363.2	25.8	435.8	685.3	6.73%	6.17%	7.1	6.0	Jul 94
Global Equity Large Cap	*Retail*	1,093.9	57.6	466.4	603.6	6.22%	5.99%	7.5	5.6	Jan 92
	Inst.	22.7	6.2	80.0	26.4	8.93%	9.33%	6.8	6.7	Feb 04
Mexico Equity	*Retail*	26.0	10.6	70.7	297.7	7.01%	9.55%	7.0	8.5	Feb 04
	Inst.	329.7	21.6	295.9	241.4	4.36%	3.85%	5.2	5.2	Apr 03
Global Fixed Income	*Retail*	1,104.1	33.9	201.9	242.3	2.63%	2.68%	5.5	4.4	Apr 05
	Inst.	211.9	24.5	126.2	427.5	6.97%	6.34%	6.0	6.0	Jan 98
Europe Equity Large Cap	*Retail*	812.6	73.8	237.7	425.9	6.71%	7.33%	7.9	6.6	Jan 92
	Inst.	349.7	19.1	298.7	638.7	2.43%	3.40%	6.0	6.6	Apr 04
Euro Fixed Income	*Retail*	1,103.6	23.7	354.7	241.2	3.83%	5.71%	8.3	5.6	Jan 92
	Inst.	234.2	13.3	252.6	459.8	11.88%	11.12%	9.6	9.5	Oct 92
US Equity Small Cap	*Retail*	456.0	43.0	280.5	580.4	10.78%	10.85%	11.0	10.2	Jan 92
	Inst.	165.0	2.1	251.2	236.9	3.86%	0.04%	5.5	5.1	Jun 07
Global Equity	*Retail*	382.6	19.5	706.9	224.1	6.68%	5.65%	9.0	5.7	Aug 95
High Yield Fixed Income	*Retail*	954.2	16.7	307.3	197.2	4.57%	3.24%	5.1	3.6	Dec 07
	Inst.	214.3	4.2	237.7	78.5	2.16%	3.09%	4.3	4.1	Apr 05
Other Fixed Income	*Retail*	651.8	10.8	202.0	59.6	2.12%	2.39%	4.6	3.3	Apr 05
	Inst.	188.1	10.8	254.7	848.5	10.89%	11.92%	9.0	7.5	Dec 92
US Equity Mid Cap	*Retail*	487.2	39.3	450.5	778.7	10.02%	11.00%	10.5	7.5	Jan 92
	Inst.	78.7	10.5	95.9	665.1	7.49%	6.10%	6.5	7.0	Dec 98
Other Europe Equity	*Retail*	547.9	68.6	281.7	209.3	9.03%	8.24%	9.1	8.0	Jan 92
	Inst.	22.2	2.4	25.7	46.4	5.59%	3.74%	7.3	10.1	Mar 04
Financials Sector Equity	*Retail*	113.0	32.8	138.8	114.9	5.29%	5.60%	8.7	6.9	Feb 01
	Inst.	73.7	1.6	131.6	11.0	3.69%	3.65%	5.6	6.6	Jan 09
Africa Fixed Income	*Retail*	90.5	1.7	90.7	22.5	5.02%	5.81%	6.3	3.4	Nov 08
Islamic Equity	*Retail*	70.5	9.5	76.4	31.0	3.65%	3.29%	6.8	7.3	Feb 07
	Inst.	69.1	1.6	56.2	24.4	4.08%	4.84%	5.3	4.4	Mar 07
Other Sector Equity	*Retail*	262.4	40.7	124.0	136.6	4.46%	2.02%	6.2	5.1	Nov 05

Table A.1 (continued)

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
	Inst.	106.9	6.8	70.7	12.3	0.66%	2.53%	5.5	4.4	Nov 07
Africa Equity	Retail	145.6	17.6	94.6	71.4	12.94%	13.52%	7.3	6.5	Apr 03
	Inst.	38.9	5.3	57.3	29.2	8.94%	7.29%	6.5	9.6	Apr 04
Technology Sector Equity	Retail	165.8	20.7	295.4	182.8	12.51%	11.56%	9.3	8.0	Jan 92
	Inst.	40.5	1.9	103.9	145.9	6.84%	9.06%	5.7	7.8	Nov 04
Energy Sector Equity	Retail	120.9	22.2	219.4	249.5	6.83%	7.39%	8.0	6.7	Jul 00
	Inst.	255.3	6.2	378.6	757.8	6.57%	5.32%	8.8	7.2	Jun 98
US Equity Large Cap Growth	Retail	551.4	21.8	1'229.6	809.9	8.65%	9.82%	11.0	8.1	Dec 92
	Inst.	194.3	4.8	415.7	863.6	7.10%	7.16%	8.3	8.0	Aug 98
US Equity Large Cap Value	Retail	381.8	21.8	1'006.4	758.0	8.62%	10.15%	10.6	6.9	Dec 92
	Inst.	403.3	16.5	573.4	1'387.8	4.92%	5.35%	9.7	8.6	Jan 92
US Fixed Income	Retail	843.9	50.5	540.4	1'540.5	4.59%	5.26%	10.6	7.6	Jan 92
	Inst.	81.6	13.3	233.5	1'000.8	5.85%	6.63%	6.8	7.7	Nov 01
Other Europe Fixed Income	Retail	263.2	7.7	327.7	185.2	4.30%	5.32%	8.2	6.9	Mar 98
	Inst.	233.1	54.6	289.9	2'019.1	8.68%	9.27%	8.4	9.8	Jan 92
US Equity Large Cap Blend	Retail	692.4	146.2	629.6	1'912.6	7.84%	8.98%	9.1	8.7	Jan 92
	Inst.	23.4	1.2	86.3	750.2	3.19%	3.91%	6.1	10.2	Jun 00
Asia Equity	Retail	111.5	6.2	142.0	758.8	4.66%	2.69%	9.1	6.7	Jan 92
	Inst.	163.0	12.9	221.8	359.3	9.28%	8.21%	7.1	6.0	Feb 04
Real Estate Sector Equity	Retail	268.8	26.8	162.3	628.0	9.86%	10.51%	8.2	6.3	Jun 96
	Inst.	108.6	8.2	303.9	222.0	3.07%	3.46%	6.5	4.8	Feb 04
Inflation Linked	Retail	160.9	15.8	284.4	391.6	4.03%	4.60%	7.8	6.2	Dec 98
	Inst.	457.1	2.5	261.1	205.3	-0.82%	4.03%	2.8	1.8	Jun 13
Emerging Markets Fixed Income	Retail	457.5	9.2	174.8	500.0	5.40%	7.32%	5.5	3.8	Mar 04
	Inst.	231.6	10.6	372.5	495.3	8.64%	8.56%	6.3	4.7	Jul 00
Emerging Markets Equity	Retail	417.3	29.5	214.4	427.7	7.69%	7.13%	7.5	5.1	May 92
	Inst.	115.1	10.1	155.3	169.1	12.09%	11.28%	6.0	7.3	Apr 03
Asia ex-Japan Equity	Retail	282.9	16.4	176.6	96.0	5.82%	7.09%	7.3	6.3	Nov 94
	Inst.	94.1	2.3	71.7	43.5	10.84%	11.45%	5.2	4.5	Apr 09
Greater China Equity	Retail	234.5	94.5	188.5	1'084.6	10.27%	10.73%	5.7	4.1	Jan 01
	Inst.	85.5	10.8	129.3	253.5	1.78%	2.31%	5.7	5.4	May 00
Japan Equity	Retail	300.9	37.6	262.0	301.1	3.84%	3.28%	6.9	5.9	Feb 98
	Inst.	59.1	8.1	324.4	542.7	5.32%	4.55%	6.1	6.4	Jan 06
UK Equity Large Cap	Retail	170.6	36.8	423.9	445.4	3.86%	3.52%	6.3	7.3	Nov 99

Table A.1 (continued)

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
	Inst.	103.3	3.3	344.8	223.4	11.47%	13.73%	4.9	4.3	May 09
Global Equity Mid/Small Cap	Retail	288.3	9.8	230.3	223.5	4.34%	6.45%	6.4	5.2	Jul 06
	Inst.	30.6	2.5	85.6	292.8	3.89%	1.81%	4.6	9.9	Apr 05
Asia Fixed Income	Retail	268.6	6.8	163.8	110.8	5.72%	4.50%	4.9	3.7	Jan 06
	Inst.	16.5	9.8	85.4	23.7	8.22%	7.30%	5.3	9.7	Mar 04
Cons. Goods & Serv. Sect. Eq.	Retail	97.0	32.0	98.2	210.0	7.41%	6.28%	7.1	7.4	Jul 00
	Inst.	51.7	6.1	275.6	98.6	2.63%	2.04%	5.5	7.1	Apr 05
Sterling Fixed Income	Retail	169.7	15.3	447.3	457.7	1.72%	2.18%	6.2	5.8	Apr 05
	Inst.	130.1	1.3	103.5	64.1	7.95%	6.91%	2.7	2.7	Dec 12
Europe Equity Mid/Small Cap	Retail	350.2	7.8	133.1	49.7	7.57%	8.75%	7.6	6.8	Jun 01
	Inst.	31.3	1.0	127.5	4.6	1.41%	-0.42%	4.1	7.6	Aug 07
Latin America Equity	Retail	86.1	10.3	140.3	764.1	8.07%	10.33%	8.1	6.3	Aug 00
	Inst.	37.5	2.0	117.1	45.8	7.65%	8.22%	6.8	8.5	Mar 04
Natural Resources Sector Equity	Retail	107.5	14.5	167.2	162.9	7.63%	6.51%	8.9	6.6	Apr 94
Brazil Equity	Retail	51.4	6.2	54.5	21.9	-0.17%	3.29%	6.5	4.9	Aug 07
	Inst.	33.7	1.0	89.1	26.0	16.92%	17.15%	1.3	2.4	Sep 08
India Equity	Retail	169.7	9.4	169.3	233.7	8.36%	7.66%	7.5	5.2	Jan 07
	Inst.	15.0	1.6	64.8	143.5	8.51%	8.42%	8.6	9.9	May 04
Utilities Sector Equity	Retail	32.6	8.1	644.6	308.8	7.38%	8.80%	11.3	8.1	Jan 92
	Inst.	32.8	4.5	77.9	62.1	9.76%	6.76%	6.5	9.7	Mar 04
Healthcare Sector Equity	Retail	173.5	22.2	415.9	342.7	7.25%	7.11%	8.3	6.1	Jun 00
	Inst.	32.5	1.0	151.4	61.7	0.89%	-0.62%	1.2	1.9	Feb 15
UK Equity Mid/Small Cap	Retail	143.2	4.6	328.5	112.0	6.56%	6.65%	6.5	5.7	Jan 06
	Inst.	4.6	4.9	9.5	8.9	7.25%	6.67%	6.0	9.5	Apr 05
Communications Sector Equity	Retail	39.5	13.7	148.8	73.3	5.73%	8.87%	7.9	9.1	Oct 01
	Inst.	83.5	5.3	33.9	29.9	6.93%	6.05%	4.5	4.4	Mar 07
Korea Equity	Retail	280.0	41.0	96.9	57.1	13.67%	13.23%	8.9	5.7	May 01
	Inst.	21.6	1.1	108.7	22.3	3.05%	3.75%	3.7	5.5	Sep 06
Asia Pacific Fixed Income	Retail	120.9	9.6	39.6	130.1	2.60%	3.34%	7.2	5.9	May 05
Thailand Equity	Retail	117.5	8.8	47.0	49.0	17.88%	17.17%	10.9	9.3	Jan 01
Other Asia Equity	Retail	78.6	3.5	89.8	30.7	1.46%	1.06%	5.2	6.3	Jan 08
Precious Metals Sector Equity	Retail	76.8	5.9	196.1	376.0	12.31%	10.76%	9.6	6.1	Jan 92
	Inst.	4.7	2.8	34.0	216.2	5.03%	11.24%	4.9	6.5	Apr 03
Canadian Equity Large Cap	Retail	159.2	13.3	368.8	379.4	8.61%	8.79%	11.1	6.6	Jan 92

Table A.1 (continued)

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
Thailand Fixed Income	Retail	82.8	1.0	126.6	165.0	3.60%	6.25%	7.2	10.8	Mar 06
South American Equity	Retail	35.4	1.2	39.0	16.9	4.51%	4.32%	9.3	6.8	Jan 06
	Inst.	14.6	1.8	33.3	367.9	7.26%	5.58%	5.1	6.5	Sep 09
Other Equity	Retail	42.5	25.3	132.9	240.9	8.27%	7.27%	8.8	6.6	Apr 96
	Inst.	6.0	5.7	31.7	16.8	7.59%	6.41%	6.6	9.0	Jan 06
Industrials Sector Equity	Retail	29.2	23.9	123.6	188.6	8.96%	10.50%	8.9	7.6	Oct 01
	Inst.	4.9	1.0	73.5	56.7	14.47%	13.70%	7.8	4.0	Dec 08
Australia & New Zealand Eq.	Retail	10.0	1.9	149.3	42.1	8.03%	8.92%	8.0	3.8	Feb 05
Canada Fixed Income	Retail	88.9	12.9	296.2	1'064.0	4.91%	5.94%	9.5	6.5	Jan 92
	Inst.	3.4	1.0	61.6	13.4	15.31%	14.31%	5.5	6.4	Apr 09
Singapore Equity	Retail	7.6	3.0	110.9	159.1	10.64%	10.25%	7.4	8.9	May 02
Canadian Eq. Mid/Small Cap	Retail	81.1	2.8	179.6	111.6	5.84%	1.42%	7.5	5.4	Apr 07
	Inst.	3.0	1.0	23.6	4.3	37.40%	29.26%	1.7	2.3	Dec 08
Taiwan Equity	Retail	135.2	9.9	55.9	363.3	10.42%	9.17%	11.1	7.4	Jul 03
Australia Fixed Income	Retail	6.7	1.2	195.8	26.5	3.24%	3.96%	2.8	3.8	Jun 10
Malaysia Fixed Income	Retail	52.9	1.0	64.5	205.8	1.81%	0.75%	6.5	9.3	Oct 07

Table C.1: Empirical rejection probabilities: Type I errors

Empirical rejection probabilities for the nominal levels $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$ for the standard (Stand), Newey–West (NW) with a bandwith of $4 \times (T/100)^{2/9}$, HC3, and our bootstrapped (Boot) significance test that evaluates the optimal block size by the method of Politis and White (2004) and the correction of Patton et al. (2009). The data was generated by sampling from the realized returns with a circular bootstrap (Boot-x) and block sizes of $x = \{1, 3, 6\}$. The simulation study includes ten US mutual funds that exhibit the entire return history from 1992 to 2016 in the Morningstar database. We sample 1,000 paths for each fund and DGP and set the alpha under the null hypothesis to the true observed alpha. We show the results for the one-factor "CAPM," three-factor "FF3," and five-factor "FF5" model with the factors "market," "size," and "value" of Fama and French (1992), and also the "momentum" of Jegadeesh and Titman (1993), and "betting against beta" factor of Frazzini and Pedersen (2014). We highlight the p-values closest to the nominal value of the test. Because the null hypothesis is true for all of the simulations, the true rejection probabilities should be equal to the nominal level of the test.

	Nominal	CAPM				FF3				FF5			
DGP	Level	*Stand*	*NW*	*HC3*	*Boot*	*Stand*	*NW*	*HC3*	*Boot*	*Stand*	*NW*	*HC3*	*Boot*
Boot-1	$\alpha = 0.10$	0.112	0.109	**0.099**	0.102	0.114	0.113	**0.099**	0.105	0.120	0.117	**0.096**	0.106
	$\alpha = 0.05$	0.061	0.059	**0.049**	0.053	0.062	0.060	**0.050**	0.056	0.064	0.061	0.046	**0.053**
	$\alpha = 0.01$	0.012	0.012	0.008	**0.010**	0.016	0.016	**0.012**	0.013	0.016	0.017	**0.011**	0.013
Boot-3	$\alpha = 0.10$	0.142	0.119	0.125	**0.111**	0.137	0.123	0.120	**0.106**	0.135	0.118	0.111	**0.105**
	$\alpha = 0.05$	0.084	0.065	0.070	**0.058**	0.078	0.068	0.066	**0.060**	0.076	0.062	0.058	**0.054**
	$\alpha = 0.01$	0.024	0.017	0.017	**0.013**	0.025	0.020	0.019	**0.015**	0.023	0.018	0.017	**0.013**
Boot-6	$\alpha = 0.10$	0.158	0.126	0.141	**0.115**	0.140	0.124	0.124	**0.112**	0.148	0.126	0.123	**0.114**
	$\alpha = 0.05$	0.098	0.071	0.085	**0.063**	0.078	0.071	0.066	**0.062**	0.088	0.071	0.068	**0.062**
	$\alpha = 0.01$	0.036	0.021	0.029	**0.017**	0.025	0.021	0.020	**0.018**	0.027	0.022	0.020	**0.015**

Table C.2: Empirical rejection probabilities: Family wise error rates (FWER)

Empirical rejection probabilities for the nominal levels $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$ and the multiple hypothesis framework of Romano and Wolf (2005a,b, 2016) controlling the FWER based on the bootstrapped (Boot-x) significance test with block sizes of $x = \{1, 3, 6, 9, 12\}$. The DGP is a circular bootstrap (Boot-x) with an optimal block size of $x = \{1, 3, 6\}$. The simulation study includes the 17 retail investment categories with a history from 1993 to 2016 from Section 2 with the investable one-factor benchmark model that is based on the value-weighted return of index funds. For each portfolio and DGP we sample 1,000 paths and set the alpha under the null hypothesis to the true observed alpha. We highlight the p-values closest to the nominal value of the test. Because for all the simulations the null hypothesis is true, the true rejection probabilities should be equal to the nominal level of the test.

DGP	Nominal Level	Boot-1	Boot-3	Boot-6	Boot-9	Boot-12
Boot-1	$\alpha = 0.10$	0.132	**0.119**	0.096	0.071	0.050
	$\alpha = 0.05$	0.066	**0.052**	0.039	0.026	0.020
	$\alpha = 0.01$	0.015	**0.009**	0.006	0.004	0.002
Boot-3	$\alpha = 0.10$	0.145	0.124	**0.098**	0.082	0.059
	$\alpha = 0.05$	0.081	0.062	**0.048**	0.040	0.030
	$\alpha = 0.01$	0.028	0.016	**0.010**	0.004	0.002
Boot-6	$\alpha = 0.10$	0.132	0.114	**0.087**	0.066	0.046
	$\alpha = 0.05$	0.073	**0.051**	0.036	0.022	0.012
	$\alpha = 0.01$	0.020	**0.008**	0.004	0.000	0.001

The Long-Only Integrated Approach to Factor Timing

Roger Rueegg

Abstract

This paper develops a novel framework to factor timing in the long-only integrated approach to style investing. A Markov switching strategy generates a timing-alpha of 0.36% per month, which solidifies the recent evidence of momentum in factor returns. The timing ability persists among different factor sets, in both developed and emerging markets, and for holding periods of up to twelve months. Also, it remains significant when we apply robust test statistics and adjust for multiple tries. Even though the strategies are subject to high turnover, trading costs can be contained with lower rebalancing frequencies.

1 Introduction

Can we successfully time the factors in a transparent long-only setting? A promising paper by Arnott et al. (2018) shows that one-month factor momentum among 51 proven factors generates high abnormal returns. Their approach to style investing involves investments in long-short single-factor portfolios. However, the clear majority of investors face short-sell constraints. Moreover, recent literature concentrated on reducing the large number of proven factors to overcome the multidimensional challenge formulated by Cochrane (2011). Hence, it is of interest as to whether a more realistic timing framework with short-selling constraints and a focus on a small set of surviving factors can generate significant timing alphas.

Consequently, we present the long-only integrated approach to factor timing. We thereby regard factors such as stock characteristics and provide a highly transparent framework that allows us to include realistic transaction costs. We show that modeling factor momentum with a Markov switching strategy that predicts the optimal factor-weights generates on average a monthly alpha of 0.36%. The alpha exists not only in the US but also in the global developed and emerging markets, among different factor sets, and for holding periods of up to twelve months. The alpha arises solely because of the timing ability because we adjust for the underlying factor exposures. In our setting, we find that the Markov switching model with two states and one month lag dominates a standard momentum strategy which invests in the optimal weights of the most recent month. However, the alpha of the one-month momentum strategy is also economically meaningful with an average alpha of 0.23% per month.

A key limitation of research on market timing is that timing ability can only be shown in retrospect. The first lesson that investment practitioners learn is that 100% of the promoted out-of-sample backtests worked in the past; however, only a few will generate high risk-adjusted returns in real-time, while the majority of the strategies will fail. As highlighted by Bailey et al. (2014) and Bailey and Lopez de Prado (2014), the reason is that the out-of-sample backtests are prone to severe selection biases. To reduce the resulting bias in our analysis, we show that the timing ability of the Markov switching strategy is robust to a battery of sensitivity tests. First, we analyze different factor sets that select from the following well-known style factors: value, profitability, investment, momentum,

low volatility, low beta, unexpected earnings, and short-term reversal. Second, we show the strategy's timing ability with three different investment universes: in the long and short-term for the US, and for the most recent investment period for the global developed and emerging markets. Since McLean and Pontiff (2016) demonstrate that the post-publication factor return is 58% lower, we pay particular attention to the most recent period when the factors were already published.

We further show that the timing ability is significant under the robust alpha test of Leippold and Rüegg (2018).[1] Moreover, the Markov switching strategy improves the annualized Sharpe ratio by 0.23 in absolute terms compared to the value-weighted market portfolio. The difference is significant even under the robust Sharpe ratio test of Ledoit and Wolf (2008). When we next compare the information ratio of the Markov switching timing strategy with the naive diversification strategy that equal-weights the factors over time, we find an average annual improvement in absolute terms of 0.53. In recent years multiple hypothesis tests, which are vital in clinical trials, have gained popularity among the finance community.[2] Consequently, we show that the timing ability also survives the multiple hypothesis adjustments of the state-of-the-art multiple hypothesis framework of Romano and Wolf (2016).[3]

A caveat of short-term timing strategies are the high turnovers. The resulting transaction costs may erase the gains from the winning bets. Based on a one-month holding period, we find an average two-sided turnover of 122% per month for the Markov switching strategy. Consequently, for markets with high transaction costs, the high costs push the one-month holding strategy into negative territory. However, we find that longer holding periods can contain the turnover and transaction costs while preserving the timing ability. Thus, accounting for higher transaction costs in the developed markets we prefer a quarterly rebalancing, and for the significant higher fees in the emerging markets a semi-annual rebalancing. From the tested combinations and after including conservative transaction costs we find that only the factor set in the US with value, profitability, investment, momentum, and low beta survives the multiple hypothesis adjustments. However, the strategies still earn an average monthly alpha of 0.29% and offer a promising improvement of the investment return.

[1] The test statistics rely on the block-resampling suggested by Lahiri (2003) for time series and on the robust Sharpe ratio test by Ledoit and Wolf (2008).

[2] See Leippold and Lohre (2012) as an early application of multiple hypothesis tests in this area, or Harvey et al. (2016) on factors.

[3] The major advantage of their test framework is that it allows for cross-dependence of the test statistics.

As a positive side-effect of our approach, we can find the ex post optimal factor weights for the integrated approach. In our long-only setting, we see that the low-risk anomaly helps to achieve higher Sharpe ratios. But, the low-risk factors are not included in the solutions with the highest information ratio. Also, we highlight that the naive diversification with equal-weights in value, profitability, and momentum attains the highest information ratio for many factor sets in the US and developed markets. Moreover, we find that the optimal factor weights change with the holding period. Momentum (WML) receives a much lower optimal weight with lower rebalancing frequencies, whereas the profitability (RMW) and investment (CMA) factor gain importance in the optimal constant weights.[4]

Our work is contributing to the puzzling evidence of cross-sectional momentum as an important investment style. The first findings of momentum go back to Levy (1967) and Jegadesh and Titman (1993) on individual stock level. Moskowitz and Grinblatt (1999) next show that stock momentum arises from industry momentum, whereas the most recent paper by Arnott et al. (2018) suggest that industry momentum may be explained by factor momentum. The literature on timing ability concentrates on both fundamental and technical predictors. An early paper by Kao and Shumaker (1999) shows a multivariate macroeconomic analysis of timing the value and growth stocks. Asness et al. (2000), Cohen et al. (2003), Asness et al. (2013), and Arnott et al. (2016) demonstrate that valuation spreads may indicate crowding in factors with a corresponding negative outlook for the affected factors. Nalbantov et al. (2006) by applying support vector regressions with technical and economic variables succeed in timing the size and value premium in the US.[5] Arshanapalli et al. (2007) use fundamental and macroeconomic factors to show that a multi-style rotation strategy based on the four style combinations with large versus small and growth versus value outperforms the buy-and-hold portfolio. Barroso and Santa-Clara (2015) find that ex ante risk adjustments double the Sharpe ratio of momentum. Besides, a recent paper by Hodges et al. (2017) demonstrates that combining different well-known predictors improves the timing ability. In contrast to the successful timing literature, Asness (2016) concludes that the factor timing performance was not convincing historically.

Based on a realistic example, we illustrate in Section 2 our novel solution on how to time the

[4]See Table 3 for the definition of the factors.

[5]Examples of technical variables are the value-growth spread or the volatility of the S&P 500, and of the economic variables are the oil price or the yield curve spread.

factors in the long-only integrated approach to style investing. Section 3 describes the empirical analysis and performs robust performance tests of our approach. In Section 4, we focus on various robustness tests. Section 5 summarizes the results.

2 The integrated approach to factor timing

There are two approaches to multi-factor investing. To illustrate the differences, we outline in Figure 1 the two frameworks. For the standard mixed approach, we first create factor portfolios F_1, F_2, and F_3. The ultimate goal in the mixed approach is now to predict the factor portfolios' returns r_{F_1}, r_{F_2}, and r_{F_3}. Based on the predicted factor returns $\hat{r}_{F_1}$, and $\hat{r}_{F_2}$, $\hat{r}_{F_3}$, we then build the portfolio by optimally mixing the individual factor portfolios. In contrast, in the integrated approach, we first build a normalized factor score for each security and factor. Next, we build the overall score that is the weighted sum of the individual factor scores. In the last step, we build the portfolio with the overall score as the decisive criteria. Thus, the ultimate goal with this approach is to predict the factor score weights that lead to the overall ranking of the securities.

[Figure 1 about here.]

For our study, we rely on the long-only integrated approach to factor investing. Because we argue that the integrated approach offers implementation advantages when it comes to timing in a realistic long-only setting. First, returns based on factors are not only the result of a metric but also the outcome of a specific portfolio construction technique. Therefore, when we apply the integrated approach to factor investing, we first fix the final portfolio construction technique and then predict the optimal building blocks that lead to the actual investment. Whereas in the standard mixed approach, it is an open issue how one constructs the portfolios once the prediction of the factor returns is available. Second, when we predict the optimal portfolio constituents, we implicitly include the interaction effects of the factors.

Let us assume that we focus on the standard US universe from the merged CRSP and Compustat database with the five factors HML, RMW, and CMA from the five-factor model of Fama and French (2015), the momentum factor WML of Jegadeesh and Titman (1993), and the low beta factor BETA

of Fama and MacBeth (1973).[6] In our analysis, we intentionally disregard the SMB factor, because we focus on the large-cap universe. In the integrated approach, we first create the factor scores $f_{m,i}$ for factors $m = 1, ..., F$ and securities $i = 1, ..., N$. To compute the factor score, we first rank the universe of stocks in ascending order according to the metrics of the five factors. We then normalize them from zero (worst) to one (best). In the next step, we build an overall score for each company, where we weight each factor score with a fixed weight w_m for each factor characteristic. To create the final portfolio, we value-weight the securities above the 70th percentile.[7]

Now, the goal of this approach is to predict the optimal weight of each factor score to finally get an overall score. To find the optimal weights over time, we first define a fixed grid of possible weights from minus one to plus one with a step size of $1/x$. We then take all the K possible combinations of these factor weights for the predefined set of factors. With five factors and a step size of 0.5 ($x = 2$) we obtain weight vectors such as $\{1, 1, 1, 1, 1\}$ or $\{0.5, 0.5, 0.5, 0.5, 0.5\}$ that lead to the same results. To remove these redundancies, we further normalize all long and short exposures within a combination to ± 1 and focus on the unique combinations. This cleaning results in an aggregated weight of the factors of plus (minus) one if there are only factors with a positive (negative) weight, or a weight of zero if there are both negative and positives weights in the combination. In our example with $F = 5$ factors and a step size of $x = 2$ this results in $K = 1,743$ combinations and a grid of weights that lie in $\pm\{1, 0.66, 0.5, 0.4, 0.33, 0.29, 0.25, 0.22, 0.2, 0.17, 0.14, 0.13, 0.11, 0\}$. The number of combinations grows exponentially with the size of the factor set. For example, for six factors and the same step size, we count 9,493 possible compositions.

With the grid of possible combinations, we compute for each period the out-of-sample returns of the different strategies. Out of the 1,743 combinations, there are several special cases. The combination with all weights equal to zero is the market strategy. Since we in this case select all the stocks without an opinion about the future impact of the individual factors. The combinations with a plus (minus) one and else zeros are the standard long (short) factor portfolios because they concentrate on a single metric only. In contrast to Fama and French (1993) we rebalance the portfolio

[6] See Table 3 for the definition of the factors.

[7] This is the standard way to build the factor portfolios as first demonstrated by Fama and French (1993). Also, it is a model-free approach that still takes the size of the companies into account, and hence reduces size biases.

in every period.[8]

Because of the discretization, we can now find the optimal weights under perfect foresight. To compute the optimal weights over time, we take the average weight of the strategies that are equal to or above the q percentile of the period. When we set $q = 100$ we obtain the special case, where we receive the weights of the best strategy only. However, we expect a higher stability in our prediction, when we average among the best strategies within a period. Thus, analogous to the portfolio construction, we concentrate on the best 30% best strategies within a holding period and set $q = 70$.[9] Figure 2 shows the optimal weights of the factors HML, RMW, CMA, WML, and BETA over time for $q \in \{70, 100\}$. By construction, the weights for $q = 100$ swing between plus and minus one. Whereas the averaging of the best 30% of the strategies in the case with $q = 70$ lead to a higher stability of the weights over time.

[Figure 2 about here.]

Further, we show in Figure 3 the log-cumulated alpha of the strategies with perfect foresight for $q \in \{70, 100\}$. The strategies with perfect foresight always invest in the optimal portfolio weights as shown in Figure 2. The alpha is computed against the market portfolio. For comparison, we also plot the alpha over time of the strategy that ex post shows the highest Sharpe ratio over time but applies only constant weights over time. For the analyzed period from June 1963 to December 2016, this is the combination that puts 40% on value as well as momentum, and 20% on profitability. We see that the strategy that invests in the average 30% best strategies over time ($q = 70$) exhibits, by construction, a lower alpha, yet it is still highly positive and stable over time. Therefore, we expect that the loss in alpha is compensated by a higher stability in the out-of-sample forecast of the timing strategies that we present in the next section. The strategy with the highest ex post Sharpe ratio over time only achieves a marginal outperformance of the perfect foresight strategies because it holds the weights constant over time. Thus, the differences between the perfect foresight strategies and the highest Sharpe ratio strategy represents the total possible outperformance when we apply timing. We now focus on the different possibilities to ex ante predict the optimal weight of the factor scores.

[8] Fama and French (1993) only rebalance in June of each year.
[9] See Section 4.4 for robustness checks for the choice of q.

The goal is to receive a proportion of this theoretical outperformance, by using only information that is known at the time of the rebalancing.

[Figure 3 about here.]

2.1 Naive diversification

The naive diversification strategy (ND) equal weights the factor scores over time. As shown in Benartzi and Thaler (2001) or DeMiguel et al. (2009) the 1/N heuristic has a long history in asset allocation and is the natural benchmark strategy to beat for the strategies presented next that time the weights. Also, it is like the integrated strategies presented in Bender and Wang (2016) or Leippold and Rueegg (2018). In our example with five factors, we consequently set the factor score weights to one third at each point in time.

2.2 One-month momentum

Motivated by the findings of Arnott et al. (2018) that find short-term persistence in the factor returns in the mixed approach, we create the one-month momentum strategy (1M) in our integrated setting. This strategy invests in the optimal weight of the previous month. Because we expect that the optimal weights of the previous month are close to the best solution of the succeeding month. While Arnott et al. (2018) regard sets of up to fifty factors and then select the best factors over time, we only concentrate on small factor sets of three to six factors. Because our intention is not to select the best factors over time, but to optimally weight an already pre-defined small set of relevant factors.[10] Consequently, once we agreed on a small set of factors, we answer the open question if factor momentum also occurs in the optimal weights over time.

2.3 Markov switching

Ever since Hamilton (1989) suggested to describe the business life cycle as a Markov switching autoregressive process, the model has been used in numerous empirical studies both in finance and

[10] Arnott et al. (2018) show that the factor momentum is also robust for a small set of factors. Hence, it is not only able to select factors, but also to time small factor models.

economics.[11] We are interested in whether a more general model that takes the interaction effects into account does a better job in predicting the optimal weights compared to the 1M strategy presented in the previous section.

Given a time-series of optimal weights $\boldsymbol{w}_t$, $t = 1, \ldots, T$, on F factors, the conditional expected weight in a regime switching model with S states are generally given by

$$\mathbb{E}\left(\boldsymbol{w}_{t+1} \mid s_t = i\right) = \sum_{j=1}^{S} p_{ij} \cdot \boldsymbol{\mu}(s_{t+1} = j), \tag{1}$$

where p_{ij} denotes the probability of moving from state i to state j in the next time step. The conditional covariance matrix of the weights $\boldsymbol{w}_{t+1}$ includes a second-order term that accounts for the moves in the conditional means as the regime changes,

$$\Sigma(s_t = i) = \sum_{j=1}^{S} p_{ij} \Sigma(s_{t+1} = j) + \sum_{j=1}^{S} p_{ii} p_{ij} \boldsymbol{\vartheta}\boldsymbol{\vartheta}^{\top}, \tag{2}$$

where $\boldsymbol{\vartheta} = \mathbb{E}\left(\boldsymbol{w}_{t+1} \mid s_{t+1} = i\right) - \mathbb{E}\left(\boldsymbol{w}_{t+1} \mid s_{t+1} = j\right)$. The distribution of $\boldsymbol{w}_{t+1}$ conditional on s_t is a mixture of S normal distributions. Therefore, the probability density function of $\boldsymbol{w}_{t+1}$ can be written as

$$f(\boldsymbol{w}_{t+1} | s_t = i) = \sum_{j=1}^{S} p_{ij} \cdot f(\boldsymbol{w}_{t+1} | s_{t+1} = j). \tag{3}$$

For the implementation of our investment strategies, we make the reasonable assumption that at the current time t, we are not entirely sure about which regime s_t is prevailing. However, we do know the structure and the relevant parameters

$$\Theta = \{\boldsymbol{\mu}_t(s_t), \Sigma_t(s_t), \boldsymbol{P}\}, \tag{4}$$

with $\boldsymbol{P}$ representing the transition matrix.

We must differentiate between three kinds of regime switching probabilities: one-step-ahead, filtered and smoothed prediction. The one-step-ahead predicted probability uses information as of

[11] For instance, the volatility feedback model of Turner et al. (1989), regime switching interest rate models as in Ang and Bekaert (1998) and regime switching VARs as in Sims and Zha (2006).

time $t-1$ to predict the regime probability in time t, whereas the filtered probability uses data as of t to estimate the likelihood in time t. Therefore, the filtered probability is more accurate compared to the one-step ahead prediction, but it suffers from a one-period look-ahead bias. Finally, there is also the smoothed estimate of regime probabilities, which means that we use all data in the entire sample to estimate the regime probability at time t. We cannot use either filtered or smoothed probabilities in the real-time forecast. Thus, for our empirical exercise, we rely only on the one-step-ahead predictions and compute the optimal weights based on the conditional mean and with the help of the toolbox created by Perlin (2015). In our standard setting, we will use the most straightforward model specification with two states ($S = 2$) and a lag of one period, and we refer to this type of strategy as the Markov switching strategy (MS). In the remainder, we apply a minimum training period of 60 months to fit the regime switching model. The training period then increases over time, as there is more history available.

2.4 Strategies over time

To give a comprehensive overview of the strategies, Table 1 shows the acronyms, definitions, and aims of the different strategies. We distinguish between three kinds of strategies. We start with the benchmark strategies that include the value-weighted market (MKT), and the ND strategy that additionally accounts for the factor returns. They both apply no factor timing and are out-of-sample. Next, there are the perfect foresight strategies with the highest Sharpe ratio (H-SR), the highest information ratio (H-IR) strategy over time, and the perfect foresight strategy (PF) that invests in each period in the best strategies as if they would be known before and thus applies timing. As the PF strategy times the factor weights, the H-SR and H-IR strategy invest with constant weights over time, but all three strategies have a severe in-sample bias. And third, the timing strategies with the 1M and MS strategies that predict the optimal factor score weights over time.

[Table 1 about here.]

Figure 4 shows the cumulated alpha of the ND, 1M, and MS strategy over time for our example with the five factors. We find that the ND benchmark strategy generates a positive alpha over the analyzed data period with respect to the MKT strategy. But, as expected, when we correct the ND

strategy for the same factor returns HML, RMW, CMA, WML, and BETA including the market return the alpha drops to 0.03% per month and is statistically insignificant from zero. In contrast, we find for the 1M and MS strategy highly consistent and significant alphas of 0.51% respectively 0.36% per month, also when we correct for their building blocks. Also, we observe that the MS and 1M strategy show a smaller drawdown in the dot-com bubble and generate stable alphas in the most recent past. Contrary, the ND strategy starts to flatten in this period. Regarding the H-SR and H-IR strategies, we find that for the Sharpe ratio it was optimal to put 28.57% for HML, RMW, and WML, as well as 14.29% for BETA what results in a Sharpe ratio of 0.62.[12]. Whereas for the relative point of view to the MKT strategy, a weight of one third on HML, RMW, and WML attains the highest information ratio of 0.66. We highlight that the optimal solution based on the information ratio disregards the CMA and BETA factor.

[Figure 4 about here.]

3 Empirical investigation

3.1 Universe

For our analysis, we consider three datasets. The first dataset is the standard universe with all non-financial NYSE, AMEX, and NASDAQ stocks from the CRSP and Compustat database. The period ranges from June 1963 to December 2017. For concerns of liquidity, we focus on the companies above the NYSE median market capitalization. It corresponds to the "Big" portfolio as described in Fama and French (1993). The second dataset is the MSCI Developed World universe, and the third the MSCI Emerging Markets universe. We have data on its constituents for the most recent twenty years from April 1998 to April 2018. Both indices invest in the most liquid large and mid-cap stocks of the developed respectively emerging markets countries. They cover approximately 85% of the free float-adjusted market capitalization within each universe. Moreover, the MSCI indices are the most commonly applied benchmark in delegated asset management. Usually, MSCI Inc. updates its constituents on a quarterly basis. While the developed and emerging market countries

[12]The MKT strategy shows a Sharpe ratio of 0.35 over the same period from June 1963 to December 2017

also change over time, but on a less frequent basis. Also, for the MSCI universes, we only include non-financials. Table 2 shows the summary statistics for the three universes. The MSCI universes hold on average 1,339 stock for the developed (DM), respectively 648 for the emerging markets (EM) universe, whereas we observe on average 813 US stocks. Since the US universe starts earlier in 1963 with a small universe of stocks, it holds on average a lower number of constituents compared to the DM universe. For the same reason, the DM universe also shows for the market capitalization a higher average market capitalization with USD million 14,234 compared to 7,033 in the US and 2,903 for the EM. The minimum market capitalization for the MSCI universes is close to zero, while it is USD million 62 for the US dataset. This difference arises because of the construction of the universes. For instance, in the US we filter each month the stocks above the NYSE median. Thus, the market capitalization never drops below this median; however, for the MSCI universes, there is only a quarterly regular rebalancing. Consequently, stocks that drop dramatically in value during a quarter are kept for a longer period compared to the US analysis. Regarding the average total excess return above the one-month Treasury bill rate we find similar average monthly total returns for the US and DM with 0.62 respectively 0.69%, the EM universe generated, on average, high returns of 1.18 for our data period.

[Table 2 about here.]

3.2 Factor sets

For the factors, we start with the five-factor model of Fama and French (2015). We disregard size since we concentrate on large-caps and the most liquid stocks around the globe only. Also, we include the momentum factor of Jegadesh and Titman (1993). For the portfolio perspective, DeMiguel et al. (2018) find that only six out of their 51 analyzed characteristics remain significant. Thus, we also include the disjunct characteristics of their analysis for our robustness check. To summarize we concentrate on the value (HML), profitability (RMW), investment (CMA), momentum (WML), unexpected earnings (SUE), low volatility (LVOL), low beta (BETA), and short-term reversal (REV) factor. Table 2 overviews the factor definitions and their acronyms. Regarding the factor sets, we focus on the Fama French factors (FF), the Fama-French including momentum as in Carhart (1997)

(FFC), the FFC model extended by low volatility and low beta (CLV, CB), and their combination including both low risk factors (CLVB). For robustness, we check our results also for the optimal factor set of DeMiguel et al. (2018) (DMNU) that includes the factors RMW, CMA, LVOL, BETA, SUE, and REV. For an overview, we show on the right of Table 2 the six factor sets from FF to DMNU and tick the included characteristics. To guarantee that we trade on information that is known at the rebalancing, we consider an additional data lag of two days for the factors in the MSCI universes. This additional lag is of importance because we trade Asian, European, and US countries where the market closes at different time zones, which creates additional implementation delays.

[Table 3 about here.]

3.3 Factor returns and optimal constant weights

For the three universes and six datasets, we discretize the possible factor score weight and calculate the different strategies over time. For the grid of factors, we set analogous to Section 2 the step size to 0.5 or $x = 2$. This choice results in 63, 325, 1,743, and 9,493 different combinations for our factor sets of three, four, five, respectively six factors. The higher we set x, the finer is the grid and the higher we expect the accuracy of our method. But due to the curse of dimensionality with a finer grid, we also obtain higher computational costs. With the choice of $x = 2$ we are in a coverable territory, where we balance the accuracy and the computational power required to calculate all the possible weight combinations over time.[13]

For the special cases with a weight of plus one or minus one on only one factor that result in the pure factor portfolios, and the combination with zero weights on all factors that results in the MKT strategy, we show the detailed summary statistics in Appendix A. For the MKT strategy, we find a highly consistent Sharpe ratio of 0.345, 0.313 and 0.325 for the US, DM, and EM universe. Also, we see higher Sharpe ratios for all factor portfolios and among all three universes with the presumably positive characteristics. Except for HML in the DM and REV in the EM universe, the assumed dominant factor-strategy led to a lower Sharpe ratio for the analyzed period of twenty years.

In Table 4 we show for each universe the ex post optimal weights in the long-only integrated

[13] See Section 4.4 for a finer grid size. E.g. for $x = 3$ we arrive at 75,985 possible combinations for six factors.

approach. In the left (right) of the table, we show the weights of the strategies with the highest Sharpe (information) ratio. First, we find that none of the factors obtain a negative weight. Thus, contrary to DeMiguel et al. (2018), who find in the long-short mixed approach that it can be optimal to take a short position in the low volatility strategy and a long position in the low beta strategy to reduce the risk, we cannot observe this kind of long-short hedging in the long-only integrated approach. Second, we find that the low-risk strategies LVOL and BETA very rarely obtain a positive weight in the highest information ratio strategies. Thus, in our setting, the low-risk strategies are only attractive from a Sharpe ratio point of view. Third, we find that the naive diversification strategy with the three factors HML, RMW, and WML offers the highest information ratio in many of the US and DM factor sets, even when we include other possible factors. Also, Fama and French (2015) mention that HML tends to be redundant when they add CMA to the factor set. However, we find that HML obtains a positive weight in all of the combinations tested, while CMA plays only a minor role in our setting and is often neglected in the factor sets. This difference can arise because we take the improved definition of HML as described in Asness and Frazzini (2013) and second rebalance the portfolio on a monthly bases.[14]

[Table 4 about here.]

3.4 Timing the factors over time

In order to extract the timing ability of the MS and 1M strategy from Section 2, we regress the return of the strategies on the multi-factor model including the market and the factors of the factor set. By controlling for the factors of the factor set, the alpha of the time series regression corresponds to the timing ability of the strategy. Since Leippold and Rueegg (2018) find that the integrated approach shows a positive sensitivity to the low-volatility strategy, and to control for other possible additional factor biases, we also correct the strategies for the eight-factor model which includes the factors MKT, HML, RMW, CMA, WML, LVOL, SUE, and REV.[15]

Table 5 shows the alpha coefficient together with the Newey-West t-values for the six factor sets of

[14] We will see in Section 4.1 that a longer holding period favors the CMA factor.

[15] We exclude the BETA factor, because LV and BETA are highly correlated with a variance inflation factor of above nine. However, the results remain the same when LV is replaced by BETA.

Table 3. We highlight significant alphas at the 95% confidence level in bold. To compute the p-values, we apply the block-resampling robust alpha test of Leippold and Rüegg (2018).[16] The results suggest that the strategies MS and 1M can significantly time the factors. With an average monthly alpha of 0.36% (0.38%) with respect to the underlying factor set (with respect to the eight-factor model), the MS strategy dominates the 1M strategy with a corresponding average alpha of 0.23% (0.23%). Also, the t-value is higher for the MS strategy with a very high value of 3.83 with respect to the same factor set of the timing strategy and even 4.22 with respect to the eight-factor model. Thus, the more sophisticated timing with the Markov switching model creates additional value over time. While the MS strategy shows significant positive alphas for every factor set for the US and DM universes, we find positive but insignificant alphas for the EM universe for the FF, FFC, CLV, and CB factor sets. We highlight that the ND strategy, in general, shows insignificant p-values. However, there are still a few significant alphas also for this naive diversification strategy. We can explain these significant values since we conduct multiple tries. Thus, we expect significant values just by luck and refer to Section 4.3 for a more detailed explanation of the impact of multiple hypothesis adjustments.[17] To conclude, we attribute the MS strategy a significant timing ability that is positive in every analysis for the three different universes and six factor sets.

[Table 5 about here.]

3.5 Performance analysis

We next analyze the return to risk differences between the MS, 1M, and ND strategy in absolute terms and from a relative perspective compared to the MKT strategy. We find in absolute terms that the Sharpe ratio of the MS strategy is on average 0.23 higher compared to the market, while the 1M strategy shows an improvement of 0.11. Regarding the Calmar ratio, where we divide the annualized return by the absolute maximum drawdown, we find an average improvement of 0.09 for the MS and 0.05 for the 1M strategy. In relative terms to the ND strategy, we find an information ratio improve-

[16] The block-resampling method shows a higher statistical power compared to other standard tests for financial time series (see Lahiri (2003) or Leippold and Rüegg (2018)). For the block size, we apply a value of five since most of the optimal block sizes lie between one and five with the method of Politis and White (2004) and Patton et al. (2009), while five results in the highest and thus most conservative p-values.

[17] With the multiple try adjustment the significant p-values start to become insignificant for the ND strategy. This behavior is expected since they apply no timing, hence the alpha must be close to zero.

ment of 0.53 for the MS and 0.10 for the 1M strategy. The relative Calmar ratio, computed as the annualized alpha divided by the maximum relative drawdown to the market, increases by 0.34 and 0.09 for the MS respectively the 1M strategy. Therefore, we conclude that the MS strategy dominates the other strategies also in the return to risk dimension and offers an economic meaningful improvement for the investor. For the interested reader we provide the detailed performance summaries for the different factor sets and universes in Appendix B and Table B.1.

To analyze the statistical significance of the Sharpe ratio difference, we show in Panel A of Table 6 the annual Sharpe ratio difference to the MKT strategy together with the Newey-West t-values. We highlight significant differences at the 95% confidence level by the robust Sharpe ratio test of Ledoit and Wolf (2008) in bold. Similar as for the alpha, we find for the MS strategy significant differences for every analysis in the US and DM universe. Still, for the EM universe, the CLV, CLVB, and DMNU offer a significant improvement for the Sharpe ratio. Also, the 1M strategy shows consistent and significant differences for the US universe; however, for the DM and EM, the significance vanishes.

In Panel B of Table 6, we also show the robust hypothesis tests with the information ratio difference of the MS and 1M strategy. For the MS strategy we find a significant improvement for the CB, and CLVB factor set in the US and the DMNU factor set in the DM and EM universes. For the 1M strategy, we only find insignificant differences, while they are mostly negative in the EM universe. In contrast, for the MS strategy, the differences are in any case positive.

We conclude that the MS strategy also dominates the return to risk comparison. After robust hypothesis tests, we find significant improvement for most of our six factor sets and three universes, while the 1M strategy shows the most substantial improvement in the US universe.

[Table 6 about here.]

4 Robustness checks

In this section, we concentrate on the dominant MS strategy and provide various robustness checks.

4.1 Longer holding periods

For our first robustness check, we increase the holding periods to 2, 3, 4, 6, and 12 months. Because of the longer holding periods, we also increase the frequency of our timing strategies. For example, when we hold the portfolio for three months, we also compute the perfect foresight strategies with a holding period of three months.[18] Moreover, the inputs to the Markov switching model are quarterly return series. In this way, we adopt the prediction for the longer holding periods because a longer holding period impacts the optimal strategies over time. Factors with a high turnover or with a weak performance for a lower rebalancing frequency obtain lower weights in the perfect foresight strategies with longer holding periods. When we compare the optimal weights of Table 4 with the optimal weights with a holding period of six months, we see that momentum, which shows a high turnover, obtains a low average weight of 6%. Whereas in the one-month analysis it contributed with on average 31% to the optimal weights. Also, SUE and REV obtain significant lower weights, when RMW and CMA obtain a much higher weight. RMW with an average weight of 39% starts to dominate for the longer holding period. The increase in the CMA factor can be explained by the decrease of HML of about the same magnitude. Thus, for the short holding periods HML with the improved definition of Asness and Frazzini (2013) seems to outperform CMA, as for longer holding periods we find as suggested by Fama and French (2015) that CMA starts to dominate.

In Figure 5 we provide the mean alpha (a), Sharpe ratio difference to the market (b), and information ratio difference to the ND strategy (c) for the holding periods from one to twelve months. Like Arnott et al. (2018) that finds the strongest results for the one-month holding and formation period, we see that for our long-term analysis in the US, the monthly alpha decreases from 0.41% for the one-month holding period to 0.20% for the two-months holding period. However, for two to twelve months it remains constant. For the DM universe, we find the same decrease of the alpha from 0.39% in the one-month holding period to 0.21% per month for the two-months holding period. Still, the alpha increases after that to 0.42% for a one-year holding period. For the EM universe, the alpha even increases from 0.28% to 0.38% and then remains constant up to a six-months holding period. For twelve months it decreases again to an average 0.29%. Since we only have 15 years of

[18] When a stock drops from the universe during the holding period, we set the weight to zero and normalize the holding weights to one.

out-of-sample period for the DM and EM universes, a more extended holding period results in only a few rebalancing over time. Thus, we must be more careful with the interpretation of these results. We also plot the 95% confidence interval of the average Newey-West standard error over the six factor sets. We observe that they lie almost in all cases above the zero line. The interested reader finds in Appendix C and Table C.1 the detailed alphas for each factor set and universe.

For the Sharpe ratio difference and the information ratio difference, we find the same patterns as for the alpha. Still, the differences are less significant compared to the alpha. However, economic meaningful with an average improvement of the Sharpe ratio of 0.14, and the information ratio of 0.46 for the holding periods from two to twelve months. Surprisingly, the Sharpe ratio difference for the EM universe is highly stable, and most of the differences are significantly different from zero at the 95% confidence interval.[19]

[Figure 5 about here.]

4.2 Transaction costs and turnover analysis

The average two-way turnover of the strategies with a holding period of one month is 122%, 132%, and 41% for the MS, 1M, and ND strategy. For the longer holding periods, the turnover decreases for the MS strategy from 63% for two to 41% for three, 32% for four, 22% for six, and 11% for twelve months. The 1M strategy shows the same speed of decrease but is always with a factor of 1.05 to 1.08 higher compared to the MS strategy. Thus, besides the better performance numbers, we also favor the MS strategy due to the lower transaction costs over time. On the other hand, the ND strategy shows only one-third of the turnover of the MS strategy for the one-month strategy. However, with the longer holding periods, the ND strategy loses the transaction costs advantage. For the six-months holding period, we find a turnover of 13%, and for twelve months of 9%, that is close to the turnover of the MS strategy. Due to the deletions and additions in the universe, also the market has an average two-way turnover of 1.6%. Thus, we also deduct small transaction costs to the MKT factor when we

[19] We compute the test statistics based on the robust test of Ledoit and Wolf (2008). We provide the detailed results of the robust hypothesis testing in Appendix C and Table C.2 for the Sharpe ratio and Table C.3 for the information ratio.

next evaluate the impact of transaction costs.[20]

In order to verify the stability of our results to the high turnover of the timing strategies, we now apply realistic trading costs to our analysis. We proceed similar as in Leippold and Rueegg (2018) but start with a 1.5% one-way transaction cost from 1963 to 1975, which is 0.5% higher due to the high turnover of the timing strategies. After the deregulation of the commission fees in 1975, we decrease the transaction costs from 1976 to 2018 at an exponential decay with a mean lifetime of twelve years. Since we find higher transaction costs and exchange fees for the exchanges outside of the US, we double the transaction costs for the DM universe. For the EM, we apply a multiplier of six. In this way we arrive at 4 Bps (US), 8 Bps (DM), and 24 Bps (EM) at the end of 2017, which lies with a small positive margin above the bid-ask spread of index funds within the same universe.[21] Also we are close to DeMiguel et al. (2018) that apply transaction costs of 1% in 1980 and 0.35% in 2002 for the largest US companies.[22].

As can be seen from Figure 6, for the US from 1968 to 2017, we find that the high transaction costs at the beginning of our analysis led to negative alphas, Sharpe ratio, and information ratio differences for holding periods up to three months. Starting from semi-annual holding periods, we find that the MS timing strategy can create value also after transaction costs. Since we see that the high transaction costs in the past prevented timing strategies from being effective, we also regard the more recent time periods starting in December 2002 for the US. We chose this starting point because the resulting analysis period of 180 months is the same as for the DM and EM universe. In this more recent analysis, we see that the one-month holding period provides the highest average alpha, Sharpe ratio and information ratio differences for the US. For the DM universe, we find the three-month, but also the twelve-months holding period to be superior. Whereas due to the high transaction costs in the emerging markets, a holding period of six to twelve months shows the highest advantage of the timing strategies. Ultimately, the overall significance diminishes with transaction costs; however, we find that only a minority of the analysis shows negative differences in the various performance

[20]The fact that we disregard transaction costs for the other long-short factors brings an advantage to the benchmark model.

[21]The bid-ask spread of index funds protects current investors from subscriptions and redemptions, and thus, is a valid approximation for the average trading costs.

[22]Since they also short securities, where transaction costs are a well-kept secret of investment banks, we are confident that our long-only transaction costs are on the conservative side.

measures. Only for the three-month holding period in the US, two-months holding period in the DM, and the one-month holding period in the EM, we find slightly negative average performance numbers. For a more detailed analysis of the statistical significance across the factor sets, we refer to the next section and the analysis including transaction costs in Panel B and Panel C of Table 7.

[Figure 6 about here.]

4.3 Multiple hypothesis tests

As outlined by Bailey et al. (2014), it is crucial to control for the number of tries for our out-of-sample tests. In the recent literature, there exist two common error rates to correct for: the false discovery rate (FDR) and the more conservative family-wise error rate (FWER).[23] The FWER approach of Romano and Wolf (2005a,b) offers first the advantage to account for the cross-dependence structure of the strategies.[24] Also, the FWER is more suitable for a lower number of hypothesis.[25] Since we obtain with six factor sets and three universes in total 18 tries per holding period, we still have a low number of hypothesis. For this reason, and to allow for cross-dependence, we apply the multiple hypothesis adjustments of Romano and Wolf (2016) to the block-bootstrapped t-statistics of the robust alpha test of Leippold and Rüegg (2018).

[Table 7 about here.]

To account for the dependence structure and to jointly sample the strategies and benchmark returns, we require connected time series. Consequently, we regard the overlapping period of the three universes starting in April 2003 and ending in December 2017. We show in Table 7 the resulting analysis with the single and multiple hypothesis adjusted p-values. We distinguish between three analysis. In Panel A we find that when we exclude transaction costs and apply a holding period of one month, the alpha remains highly significant, also when we correct for multiple tries. However, when we include transaction costs, the significance vanishes with the adjustment for multiple tries.

[23] See, e.g., Romano et al. (2007) or Bajgrowicz and Scaillet (2012) for a discussion of the differences in the two metrics.

[24] We find an average cross-correlation of 0.12 among the six factor sets in the three universes.

[25] Bajgrowicz and Scaillet (2012) argue that for a high number of strategies, it is favorable to control for the less restrictive FDR. Because the chance to miss an outperforming strategy is worse in a well-diversified approach compared to invest in a few false discoveries.

Since we in the previous section find, that the high transaction costs eliminate the gains in markets with higher transaction costs, it follows naturally to apply a less frequent rebalancing for markets with higher transaction costs. Thus, when we regard the analysis in Panel C with a monthly rebalancing for the US, a quarterly rebalancing for the DM with higher transaction costs, and a six-months holding period for the EM with the highest fees, we see that the alphas after transaction costs are economically meaningful with 0.29% per month, but also above the 10% significance level. CB in the US universe, that shows a monthly alpha of 0.41%, is the only factor set that remains significant at the 90% confidence level when we correct for transaction costs and multiple tries.

4.4 Further comments on the parameter sensitivity

In our paper, the focus of attention is on a low number of parameters with only a few optimizations over time. In this way, we keep the computational flexibility to compute the parameter sensitivities. We first had to define the step size $1/x$ for the discretization of the weights grid. The choice of $x = 2$ delivers highly robust results over all tested universes and combinations. When we increase x to three, we find the same average monthly alpha of 0.36% per month for the one-month holding period as in Table 5. The second parameter q that defines the lower percentile of the best strategies over time is set to 70%. Thus, we not only invest in the 30% best stocks for our portfolio construction but also predict the average weight of the 30% best strategies over time. Also, in the analysis with $q = 80$, we find the same average alpha of 0.36% for the MS strategy as in the standard setting. When we set the parameter higher to the extreme value of $q = 100$, where we only predict the weight of the best strategy, the prediction is more prone to estimation errors, which leads to less robust and on average inferior results.

In tests not reported, we also compute the sector-adjusted performance of our approach. We find that the average alpha for the multiple tries analysis of the previous section is insignificantly below the reported alphas. For instance, we find in the industry-adjusted comparison within the three universes a monthly alpha of 0.31% compared to the 0.34% reported in Panel A of Table 7.

When we intend to reduce transaction costs, we can also apply other portfolio constructions with restrictions on the turnover. However, when we restricted the turnover over time in our analysis, the

limitations on trading prevented the timing strategies to place their bets over time. Thus, we find it optimal to increase the prediction and holding periods, instead of restricting the turnover of the strategies to adjust for transaction costs.

We also tested other portfolio constructions techniques. However, since the beauty of our approach is to fix the final portfolio construction and then solve for the optimal weights, an alternative portfolio construction technique removes this direct link. Thus, we leave further improvements on the portfolio constructions such as implementing a risk model to the practitioners and concentrate on the standard model-free value-weights of the top 30% scored stocks.

5 Conclusion

Yes, we can successfully time the factors in a realistic long-only setting. We present a novel framework to factor timing that relies on the long-only integrated approach to style investing. In contrast to predicting the factor returns of long-short portfolios, we forecast the optimal weights of the stocks' factor-characteristics. We show that a Markov switching model with two states and based on the prior months' results generates an average alpha of 0.36% per month. The strategy dominates the performance comparison when we compare it to a strategy that invests in the optimal weights of the past month. We also find a significantly higher Sharpe ratio relative to the market and an economic meaningful information ratio improvement of 0.53 in absolute terms compared to the naive diversification strategy that equal-weights the factors. Our findings support Arnott et al. (2018), in that we provide further evidence for the high returns of momentum strategies applied to factor investing.

The timing ability is present among different factor sets including the most common investment styles. The results are also valid for the highest capitalized stocks within the US, developed and emerging markets universe. When we increase the holding and forecasting period to up to twelve months, the timing ability weakens, but is still meaningful and significant. The significance is robust to the state-of-the art block resampling method of Ledoit and Wolf (2008) and Leippold and Rüegg (2018). This is even true, when we adjust for multiple tries as suggested by Bailey et al. (2014).

One limitation of the short-term timing strategies are the high turnovers. In markets such as

the emerging markets or the more distant past, where transaction costs were high, monthly holding periods were not the optimal choice. When we apply longer holding periods for markets with higher transaction costs, we find that the Markov switching strategy generates an economical meaningful alpha of 0.29% per month over the most recent past.

To conclude, there is still the open question, whether beating the market with a successful factor selection, or only creating additional value once one agreed on a factor set is a source of alpha. We show that a Markov switching strategy can generate alpha, where the decisions are only based on the historical returns of the factor characteristics. Hence, we find evidence that the market prices of risk only adjusted slowly over time and that it was a source of alpha in retrospect.

References

Ang, Andrew, and Geert Bekaert, 1998, Regime switches in interest rates, *Journal of Business & Economic Statistics* 20, 163–182.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259–299.

Arnott, Robert D., Noah Beck, Vitali Kalesnik, and John West, 2016, How can "smart beta" go horribly wrong?, Technical report, Research Affiliates.

Arnott, Robert D., Mark Clements, Vitali Kalesnik, and Juhani T. Linnainmaa, 2018, Factor momentum, Available at SSRN 3116974.

Arshanapalli, Bala G., Lorne N. Switzer, and Karim Panju, 2007, Equity-style timing: A multi-style rotation model for the russell large-cap and small-cap growth and value style indexes, *Journal of Asset Management* 8, 9–23.

Asness, Clifford S., 2016, The siren song of factor timing aka "smart beta timing" aka "style timing", *Journal of Portfolio Management* 42, 1–6.

Asness, Clifford S., and Andrea Frazzini, 2013, The devil in HML's details, *Journal of Portfolio Management* 39, 49–68.

Asness, Clifford S., Jacques A. Friedman, Robert J. Krail, and John M. Liew, 2000, Style timing: Value versus growth, *Journal of Portfolio Management* 26, 50–60.

Asness, Clifford S., Tobias J. Moskowitz, and Lasse Heje Pedersen, 2013, Value and momentum everywhere, *Journal of Finance* 68, 929–985.

Bailey, David, and Marcos Lopez de Prado, 2014, The deflated sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality, *Journal of Portfolio Managment* 40, 94–107.

Bailey, David H., Jonathan M. Borwein, Marcos L. de Prado, and Qiji J. Zhu, 2014, Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance, *Notices of the American Mathematical Society* 61, 458–471.

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical trading revisited: False discoveries, persistence tests, and transaction costs, *Journal of Financial Economics* 106, 473–491.

Barroso, Pedro, and Pedro Santa-Clara, 2015, Momentum has its moments, *Journal of Financial Economics* 116, 111–120.

Benartzi, Shlomo, and Richard H. Thaler, 2001, Naive diversification strategies in defined contribution saving plans, *American Economic Review* 91, 79–98.

Bender, Jennifer, and Taie Wang, 2016, Can the whole be more than the sum of the parts? Bottom-up versus top-down multifactor portfolio construction, *Journal of Portfolio Management* 42, 39–50.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Cochrane, John H., 2011, Presidential address: Discount rates, *Journal of Finance* 66, 1047–1108.

Cohen, Randolph B., Christopher Polk, and Tuomo Vuolteenaho, 2003, The value spread, *Journal of Finance* 58, 609–641.

Cooper, Michael J., Huseyin Gulen, and Michael J. Schill, 2008, Asset growth and the cross-section of stock returns, *Journal of Finance* 63, 1609–1651.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2009, Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?, *Review of Financial Studies* 22, 1915–1953.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J. Nogales, and Raman Uppal, 2018, A portfolio perspective on the multitude of firm characteristics, CEPR discussion paper DP12417, Centre for Economic Policy Research.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–36.

Hamilton, James D., 1989, A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* 57, 357–384.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, ...and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Hodges, Philip, Ked Hogan, Justin R. Peterson, and Andrew Ang, 2017, Factor timing with cross-sectional and time-series predictors, *Journal of Portfolio Management* 44, 30–43.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Jegadesh, Narasimhan, 1990, Evidence of predictable behavior of security returns, *Journal of Finance* 45, 881–898.

Jegadesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Kao, Duen-Li, and Robert D. Shumaker, 1999, Equity style timing, *Financial Analysts Journal* 55, 37–48.

Lahiri, Soumendra N., 2003, *Resampling methods for dependent data* (Springer–Verlag).

Ledoit, Olivier, and Michael Wolf, 2008, Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Leippold, Markus, and Harald Lohre, 2012, Data snooping and the global accrual anomaly, *Applied Financial Economics* 22, 509–535.

Leippold, Markus, and Roger Rueegg, 2018, The mixed vs the integrated approach to style investing: Much ado about nothing?, *European Financial Management* 24, 829–855.

Leippold, Markus, and Roger Rüegg, 2018, Is active investing a zero-sum game?, Available at SSRN 3107904.

Levy, Robert A., 1967, Relative strength as a criterion for investment selection, *Journal of Finance* 22, 595–610.

McLean, David R., and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability?, *Journal of Finance* 71, 5–32.

Moskowitz, Tobias J., and Mark Grinblatt, 1999, Do industries explain momentum?, *Journal of Finance* 54, 1249–1290.

Nalbantov, Georgi, Rob Bauer, and Ida Sprinkhuizen-Kuyper, 2006, Equity style timing using support vector regressions, *Applied Financial Economics* 16, 1095–1111.

Novy-Marx, Robert, 2013, The other side of value: The gross profitability premium, *Journal of Financial Economics* 108, 1–28.

Patton, Andrew, Dimitris N. Politis, and Halbert White, 2009, Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White, *Econometric Reviews* 28, 372–375.

Perlin, Marcelo, 2015, MS_Regress-the MATLAB package for Markov regime switching models, Available at SSRN 1714016.

Politis, Dimitris N., and Halbert White, 2004, Automatic block-length aelection for the dependent bootstrap, *Econometric Reviews* 23, 53–70.

Rendleman, Richard, Charles P. Jones, and Henry A. Latane, 1982, Empirical anomalies based on unexpected earnings and the importance of risk adjustments, *Journal of Financial Economics* 10, 269–287.

Romano, Joseph P., and Michael Wolf, 2005a, Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* 100, 94–108.

Romano, Joseph P., and Michael Wolf, 2005b, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P., and Michael Wolf, 2016, Efficient computation of adjusted p-values for resampling-based stepdown multiple testing, *Statistics & Probability Letters* 113, 38–40.

Romano, Joseph P., Michael Wolf, et al., 2007, Control of generalized error rates in multiple testing, *The Annals of Statistics* 35, 1378–1408.

Rosenberg, Barr, Kenneth Reid, and Ronald Lanstein, 1985, Persuasive evidence of market inefficiency, *Journal of Portfolio Management* 11, 9–16.

Sims, Christopher A., and Tao A. Zha, 2006, Were there regime switches in us monetary policy, *American Economic Review* 96, 54–81.

Turner, Christopher M., Richard Startz, and Charles R. Nelson, 1989, A Markov model of heteroskedasticity, risk and learning in the stock market, *Journal of Financial Economics* 25, 3–22.

A Factor returns

In this appendix, we provide the summary statistics for the factors presented in Table 3. Contrary to Fama and French (1993), we rebalance the portfolios every month. However, we apply the same portfolio construction, when we show the value-weighted return of the best and bottom 30% of the stocks for the top (left column) and bottom (right column) portfolio in Table A.1. We find for every factor in the three universes a higher Sharpe ratio for the top compared to the bottom portfolio. The only exception builds the HML factor in the DM and the REV factor in the EM universe. When we compare the monthly mean return of HML from June 1926 to April 1998 with the period from April 1998 to April 2018 for the entire US universe including small caps, we find a mean return of monthly 0.45% relative to 0.13%.[26] Thus, we explain the low performance of value in the DM universe with the in general low returns of value for the same period in other universes. Regarding the REV factor in the EM universe, we observe similar mean returns in the period before and after 1998 for other universes. Thus, the weak performance only holds for this period and market.

[Table A.1 about here.]

B Detailed performance analysis

This appendix shows the detailed performance analysis of the MS, 1M, and ND strategy in absolute terms and relative to the MKT strategy. We provide in Table B.1 the summary statistics for the three universes from Table 2 and the six factor sets from Table 3. We find that the MS strategy offers the highest Sharpe and Calmar ratios in a clear majority of the comparisons. Except for the FF factor set in the US, the MS strategy always offers the highest return among the three strategies.

[Table B.1 about here.]

[26] The long-term factor returns are retrieved from the homepage of Kenneth French.

C Detailed results of longer holding period analysis

Table C.1 presents the detailed analysis of the monthly alphas of the MS timing strategy when we apply longer holding periods to the strategies. The alpha is relative to the factors of the underlying factor sets. Thus, the alpha represents the pure timing ability of the MS strategy.

[Table C.1 about here.]

Table C.2 reports the detailed analysis of the annualized Sharpe ratio differences of the MS timing strategy and the MKT strategy when we apply longer holding periods to the strategies.

[Table C.2 about here.]

Table C.2 gives a detailed analysis of the annualized information ratio differences of the MS timing strategy relative to the ND strategy when we apply longer holding periods to the strategies. The ND strategy is the natural equal-weighted benchmark when we apply no timing on the factor characteristics.

[Table C.3 about here.]

Figure 1: Description of the difference in the long-only mixed and integrated approach to factor timing

The mixed approach to factor timing that is standard in the literature compared to our long-only integrated approach to factor timing. We illustrate the different steps that lead to the optimal long-only portfolio with $N = 10$ stocks and three factors.

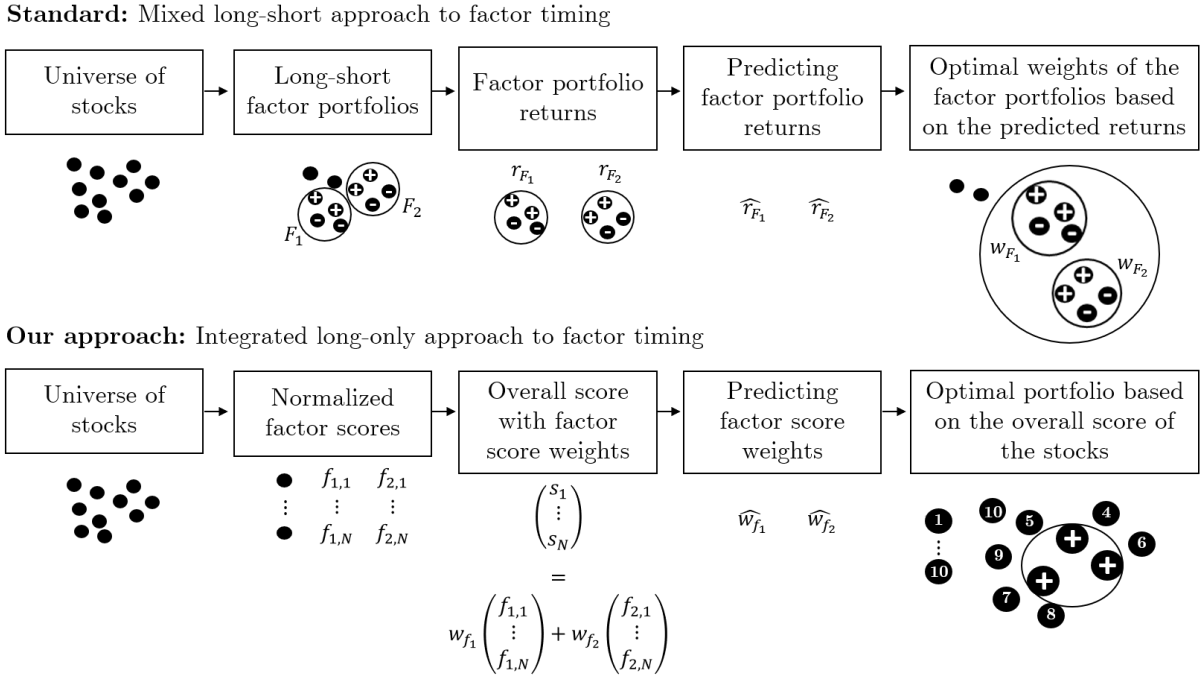


Figure 2: Optimal weights over time

Optimal weight over time of the best strategy ($q = 100$) with the dotted line, and the average weight of the 30% best strategies ($q = 70$) with the dashed line. We show the ex post optimal weight of the three factors value (HML), profitability (RMW), investment (CMA), momentum (WML), and low beta (BETA) in the US universe from 1963 to 2017.

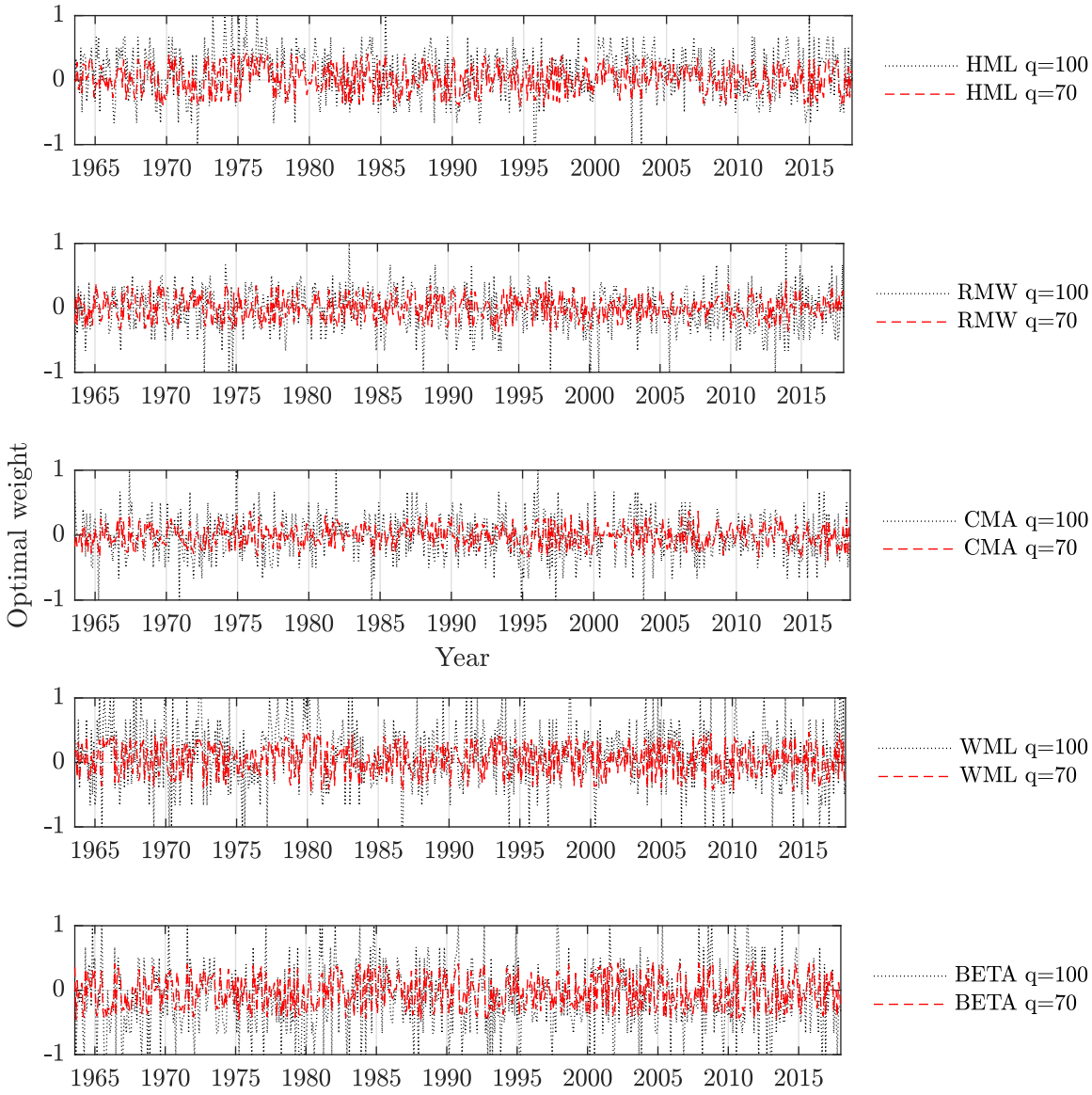


Figure 3: Alpha over time

The cumulative excess log-return to the market of the perfect foresight strategies that always invest in the best strategy (PF $q = 100$) with the dashed line, and in the average weights of the strategies above the 70th percentile (PF $q = 70$) with the solid line. For comparison, we also plot with the dash-dotted line the cumulative log-alpha of the strategy that applies the constant weights over time that lead to the highest Sharpe ratio (H-SR).

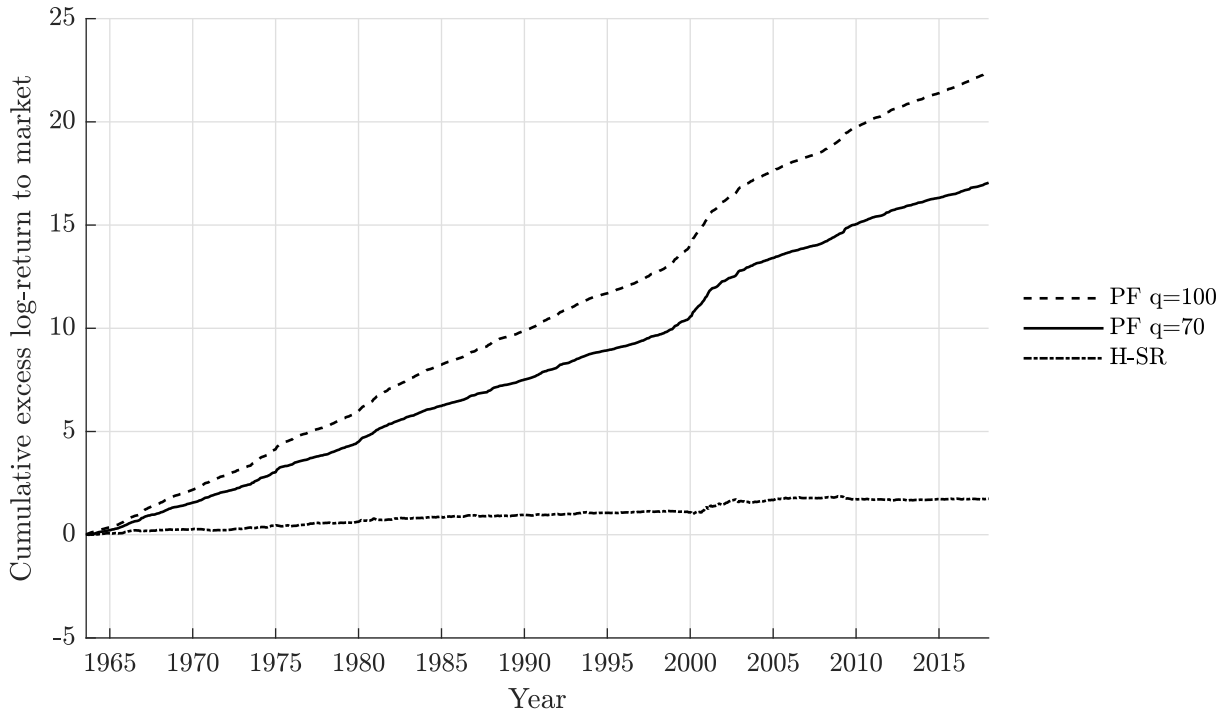


Figure 4: Cumulated alpha of the timing strategies for value and momentum

The cumulative excess log-return to the value-weighted market return (MKT) of the Markov switching strategy (MS) with the solid line, the one-month momentum strategy (1M) with the dotted line, and the naive diversification strategy (ND) with the dash-dotted line. The analyzed period is from June 1968 to December 2016.

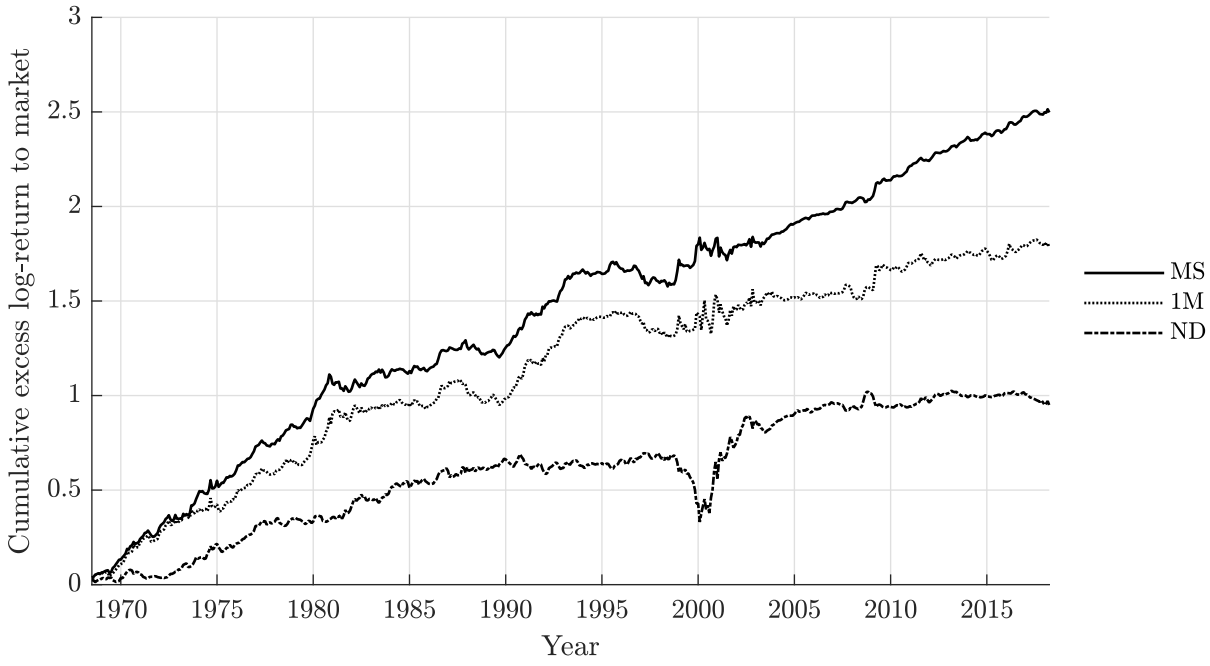


Figure 5: Alpha, Sharpe ratio, and information ratio difference

Averages among the six factor sets from Table 3 of the alpha relative to the strategies' factor set (a), the Sharpe ratio difference to the market (b), and the information ratio difference to the naive diversification (ND) strategy (c). We distinguish between the three universes from Table 2 with the US from June 1968 to December 2016, and the MSCI Developed Markets (DM) as well as the MSCI Emerging Markets (EM) from April 1998 to April 2018. The error bounds are the 95% confidence bounds based on the mean Newey-West standard error.

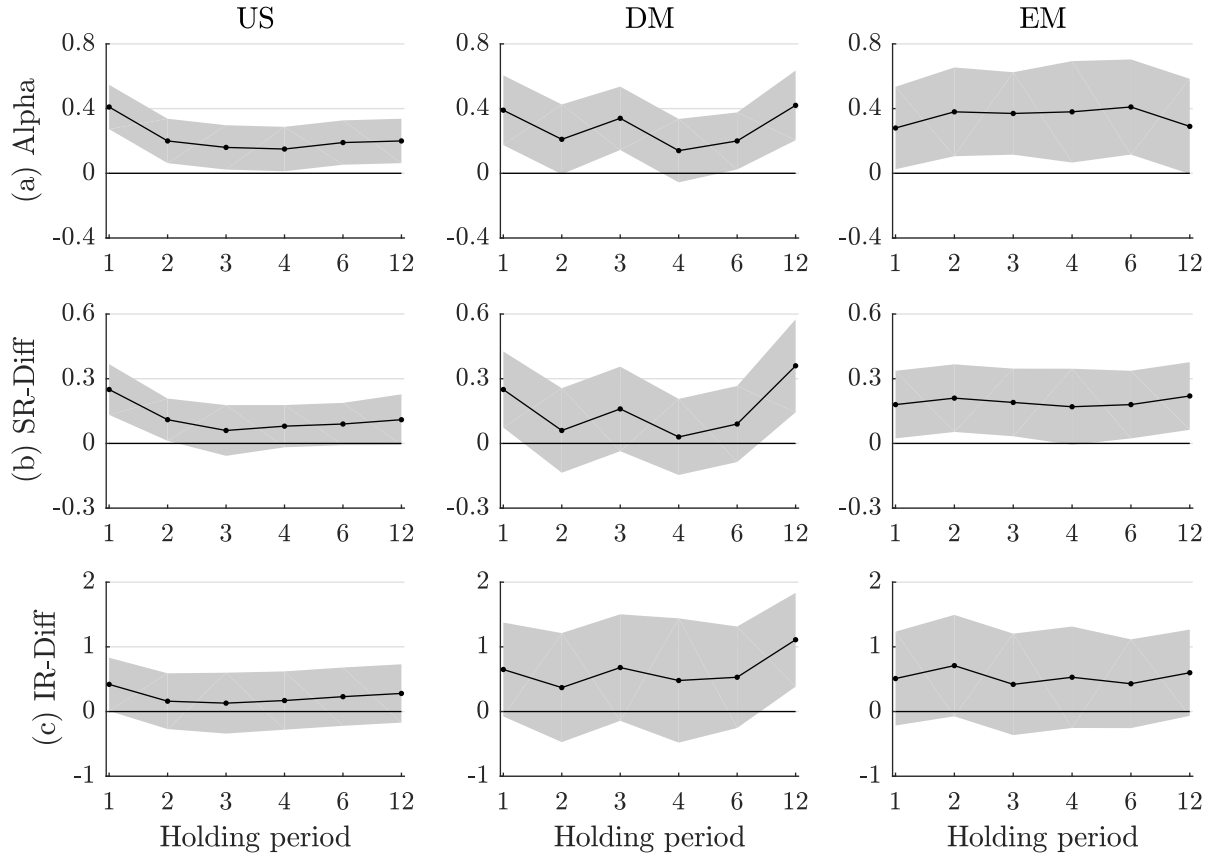


Figure 6: Transaction costs analysis

Averages among the six factor sets from Table 3 for the alpha relative to the strategies' factor set (a), the Sharpe ratio difference to the market (b), and the information ratio difference to the naive diversification (ND) strategy (c). The solid line and confidence bounds include transaction costs, while the dashed line does not include transaction costs. We distinguish between the three universes from Table 2 with the US from June 1968 to December 2016, and the DM as well as the EM from April 1998 to April 2018. For the US we show the analysis for the period starting in June 1968 (68-17) and in December 2002 (03-17). The error bounds are the 95% confidence bounds based on the average Newey-West standard error.

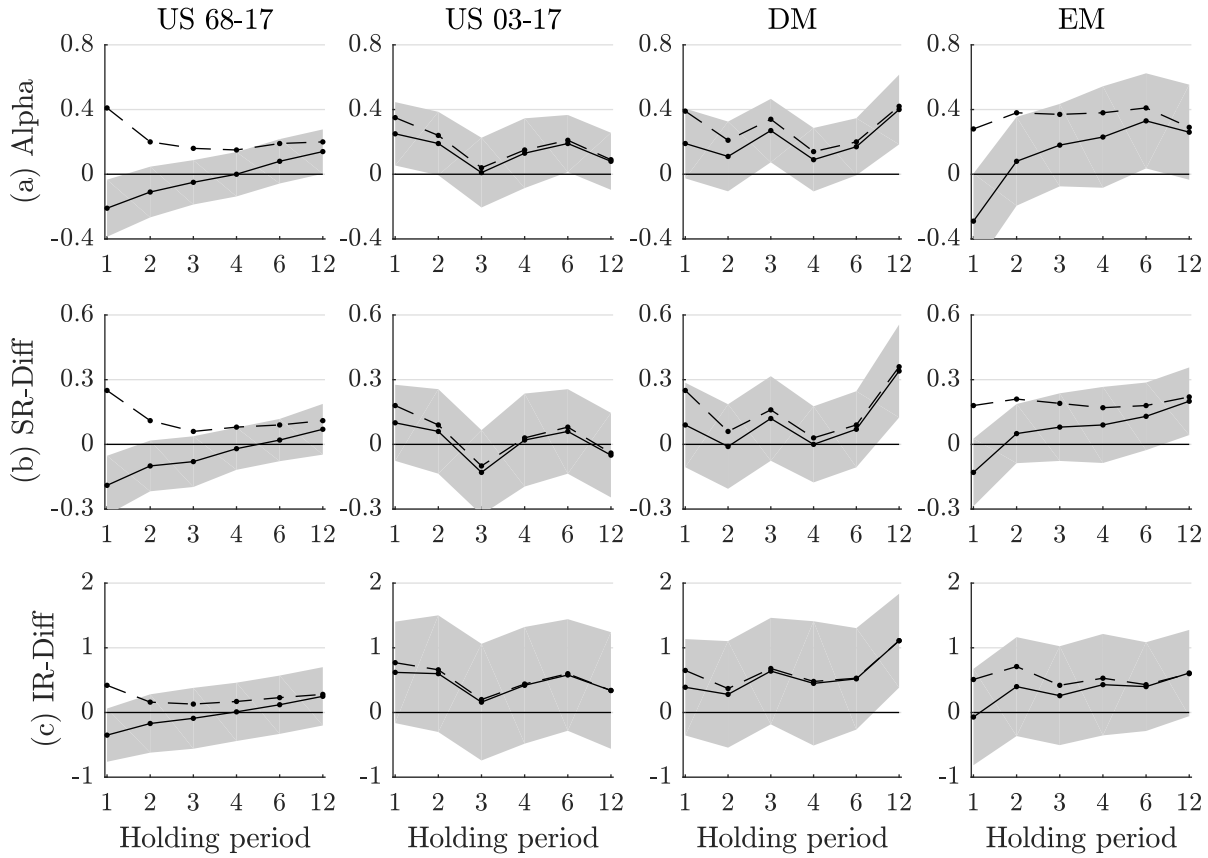


Table 1: Overview of the strategies

The acronym, the underlying model, the definition, and the aim of the strategies. The last two columns show if the strategy applies factor timing and if the strategy is out-of-sample.

Acronym	Model	Definition	Aim	Factor timing	Out-of sample
MKT	Capital Asset Pricing Model	Market-capitalization weighted portfolio	Benchmark	×	✓
Factor strategies:		*Long-only integrated approach based on the 30% highest scored stocks when the score is built with ...*			
ND	Naive Diversification	... equal weights for each factor	Benchmark	×	✓
H-SR	Perfect Foresight	... the constant weights over time that show the highest Sharpe ratio over time.	Illustrative	×	×
H-IR	Perfect Foresight	... the constant weights over time that show the highest information ratio over time.	Illustrative	×	×
PF	Perfect Foresight	... the time-varying weights that show the highest return for the actual period.	Illustrative	✓	×
1M	Momentum	... the time-varying weights that show the highest return in the most recent month.	Timing	✓	✓
MS	Markov Switching	... the prediction of the factor weights by a two regime Markov switching model.	Timing	✓	✓

Table 2: Summary statistics for the three universes

The average, minimum, and maximum number of stocks, as well as the market capitalization, and total excess return in percentage above the one-month Treasury bill rate. Also, it shows the start and end date of the analyzed data period. The universes are all NYSE, AMEX, and NASDAQ stocks above the NYSE median for the US, and the MSCI Developed Markets (DM) and MSCI Emerging Markets (EM) universe of MSCI Inc. The market cap for the US is the CRSP market cap and for the MSCI universes the free-float adjusted market cap by MSCI Inc.

	Number			Market Cap M USD			Monthly Return USD			Period	
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Start Date	End Date
US	813	587	1,550	7,033	62	882,332	0.62	-98.30	299.74	June 1963	Dec 2017
DM	1,339	1,093	3,844	14,234	0	924,421	0.69	-100.00	386.48	April 1998	April 2018
EM	648	514	2,337	2,903	0	343,776	1.18	-90.97	453.21	April 1998	April 2018

Table 3: Factor definitions and authors

Factors' names in the first, definition in the second, authors with the year of publication in the third, and acronyms in the fourth column. In headers five to ten to the right of the table, we show the acronyms of the factor sets and indicate with a ✓ the included characteristics.

Factor	Definition	Author	Acronym	FF	FFC	CLV	CB	CLVB	DMNU
Value	Devil definition as defined in Asness and Frazzini (2013) with the lagged book equity divided by the market value of the last month	Rosenberg et al. (1985)	HML	✓	✓	✓	✓	✓	
Profitability	Operating profitability as defined by sales minus cost of goods sold divided by book equity	Novy-Marx (2013)	RMW	✓	✓	✓	✓	✓	✓
Investment	Inverse annual total book asset growth	Cooper et al. (2008)	CMA	✓	✓	✓	✓	✓	✓
Momentum	Total return in US dollar of the past twelve months excluding the most recent month	Jegadesh (1990)	WML		✓	✓	✓	✓	
Low volatility	Inverse volatility for the most recent 36 months for the US dataset, and for the most recent 22 days for the MSCI datasets	Ang et al. (2006)	LVOL			✓		✓	✓
Low beta	Inverse estimated beta from monthly returns for the most recent 36 months against the equal weighted market return for the US dataset, and from weekly returns for the most recent 52 weeks against the MSCI World or Emerging Markets Index for the MSCI datasets	Fama and MacBeth (1973)	BETA				✓	✓	✓
Unexpected Earnings	Unexpected earnings approximated by the difference in the yearly change in the operating profitability from the most recent to the past year	Rendleman et al. (1982)	SUE						✓
Reversal	Inverse total return in US dollar of the past month	Jegadesh (1990)	REV						✓

Table 4: Optimal constant weights over time

Optimal weights in the long-only integrated approach for the US from 1963 to 2016 (US), the DM from 1998 to 2016, and the EM from 1998 to 2016. In the first column, we show the Sharpe and information ratio for the highest Sharpe ratio (H-SR) and highest information ratio strategy (H-IR). The rows represent the factor sets, and the sideway columns the factors from Table 3. The numbers are in percentage.

Panel US: June 1963 to December 2017

	H-SR									H-IR								
	SR	HML	RMW	CMA	WML	LVOL	BETA	SUE	REV	IR	HML	RMW	CMA	WML	LVOL	BETA	SUE	REV
FF	51.8	40	40	20	–	–	–	–	–	40.9	33	67	0	–	–	–	–	–
C	58.6	40	20	0	40	–	–	–	–	65.5	33	33	0	33	–	–	–	–
CLV	58.6	40	20	0	40	0	–	–	–	65.5	33	33	0	33	0	–	–	–
CBETA	61.8	29	29	0	29	–	14	–	–	65.5	33	33	0	33	–	0	–	–
CLVBETA	61.8	29	29	0	29	0	14	–	–	65.5	33	33	0	33	0	0	–	–
DMNU	54.5	–	14	29	–	0	29	0	29	44.7	–	33	0	–	0	0	33	33

Panel DM: April 1998 to April 2018

	H-SR									H-IR								
	SR	HML	RMW	CMA	WML	LVOL	BETA	SUE	REV	IR	HML	RMW	CMA	WML	LVOL	BETA	SUE	REV
FF	52.1	33	33	33	–	–	–	–	–	52.2	40	40	20	–	–	–	–	–
C	55.0	33	33	17	17	–	–	–	–	64.1	33	33	0	33	–	–	–	–
CLV	62.9	14	14	14	29	29	–	–	–	64.1	33	33	0	33	0	–	–	–
CBETA	64.2	22	22	22	11	–	22	–	–	64.1	33	33	0	33	–	0	–	–
CLVBETA	66.8	20	20	10	20	10	20	–	–	64.1	33	33	0	33	0	0	–	–
DMNU	62.3	–	29	0	–	14	29	0	29	64.2	–	50	0	–	0	0	25	25

Panel EM: April 1998 to April 2018

	H-SR									H-IR								
	SR	HML	RMW	CMA	WML	LVOL	BETA	SUE	REV	IR	HML	RMW	CMA	WML	LVOL	BETA	SUE	REV
FF	54.8	17	33	17	33	–	–	–	–	55.4	50	25	0	25	–	–	–	–
C	59.9	17	17	0	33	33	–	–	–	73.1	25	0	0	50	25	–	–	–
CLV	61.0	17	17	0	33	–	33	–	–	82.7	33	17	17	17	–	17	–	–
CBETA	68.5	29	14	0	29	14	14	–	–	73.5	25	0	0	50	25	0	–	–
CLVBETA	71.6	–	33	0	–	17	33	17	0	82.7	–	50	0	–	0	0	25	25
DMNU	67.8	21	27	7	32	21	27	17	0	65.3	33	26	3	35	17	6	25	25

Table 5: Out-of-sample timing ability

Monthly alphas in percentage of the Markov switching (MS), the momentum (1M) and the naive diversification (ND) strategies together with the Newey West t-value and a bandwidth of $4 * (T/100)^{(2/9)}$ in italics for the six factor models from FF to DMNU of Table 3 and the three universes from Table 2 with the US from June 1968 to December 2017, and DM as well as EM from April 2003 to April 2018. For the Markov switching model, we put an initial learning period of 5 years. We show in Panel A the alphas relative to the factors in the factor model including MKT and in Panel B the alphas relative to the factor model including the uncorrelated factors MKT, HML, RMW, CMA, WML, LVOL, SUE, and REV. We highlight significant alphas below the 5% significance level by the robust alpha test of Leippold and Rüegg (2018) in bold. In the bottom of the table, we provide the Panels' averages.

Panel A: Relative to the strategies' underlying factor model

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
US	**0.33**	**0.36**	0.03	**0.28**	**0.30**	0.00	**0.41**	**0.32**	0.02	**0.51**	**0.36**	0.03	**0.51**	**0.32**	0.02	**0.45**	**0.33**	0.03
	5.32	*5.97*	*1.04*	*4.66*	*4.13*	*0.07*	*5.24*	*3.84*	*0.76*	*7.52*	*4.41*	*1.28*	*6.44*	*3.66*	*0.72*	*5.65*	*4.93*	*1.60*
DM	**0.28**	**0.24**	0.07	**0.35**	0.25	−.02	**0.41**	0.30	**0.08**	**0.34**	0.28	0.06	**0.38**	0.30	0.05	**0.58**	**0.31**	0.07
	3.85	*3.00*	*1.72*	*3.22*	*1.84*	*0.51*	*3.99*	*2.01*	*2.42*	*3.09*	*1.77*	*1.85*	*3.01*	*1.91*	*1.91*	*4.62*	*3.36*	*1.93*
EM	0.13	0.12	0.04	0.11	0.07	−.04	0.27	0.10	0.00	0.25	0.11	**0.19**	**0.50**	0.12	**0.18**	**0.42**	−.15	0.06
	1.26	*1.13*	*0.59*	*0.89*	*0.58*	*0.78*	*1.85*	*0.68*	*0.05*	*1.61*	*0.75*	*3.70*	*3.44*	*0.68*	*2.86*	*3.27*	*1.01*	*1.03*

Panel B: Relative to the eight-factors model

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
US	**0.39**	**0.42**	0.03	**0.32**	**0.38**	0.00	**0.46**	**0.43**	0.02	**0.56**	**0.46**	0.04	**0.57**	**0.43**	0.03	**0.48**	**0.37**	0.04
	6.98	*8.10*	*1.23*	*5.32*	*6.29*	*0.13*	*5.84*	*6.53*	*0.78*	*8.12*	*7.14*	*1.37*	*7.16*	*6.57*	*1.04*	*5.95*	*5.38*	*1.90*
DM	**0.28**	**0.24**	0.06	**0.35**	**0.25**	−.01	**0.41**	**0.29**	**0.08**	**0.34**	**0.28**	0.09	**0.39**	**0.32**	0.08	**0.57**	**0.30**	0.09
	4.39	*3.44*	*1.66*	*3.89*	*2.56*	*0.46*	*4.03*	*2.67*	*2.36*	*4.10*	*2.83*	*1.93*	*3.30*	*2.91*	*1.84*	*4.52*	*3.27*	*2.12*
EM	0.11	0.06	0.06	0.14	−.01	−.04	0.25	−.01	0.01	0.22	0.02	**0.25**	**0.49**	0.00	**0.23**	**0.43**	−.16	0.12
	1.07	*0.56*	*0.92*	*1.13*	*0.08*	*0.78*	*1.76*	*0.08*	*0.13*	*1.60*	*0.14*	*3.59*	*3.50*	*0.00*	*2.95*	*3.25*	*1.01*	*1.56*

Averages:

Panel A	MS	1M	ND
Alpha	0.36	0.23	0.05
t-value	*3.83*	*2.54*	*1.38*

Panel B	MS	1M	ND
Alpha	0.38	0.23	0.07
t-value	*4.22*	*3.31*	*1.49*

Table 6: Out-of-sample robust Sharpe ratio and information ratio test

Annual Sharpe ratio difference to the MKT strategy (Panel A), and the annual information ratio difference to the ND strategy (Panel B) together with the Newey West t-statistics and a bandwith of $4 * (T/100)^(2/9)$ in italics of the MS, 1M and ND strategies from Table 1. We distinguish between the three universes US, DM, and EM from Table 2 and the six factor sets FF to DMNU from Table 3. We highlight in bold significant differences with the robust Sharpe ratio test of Ledoit and Wolf (2008) below the 5% significance level. In the bottom of the table, we provide the Panels' averages.

Panel A: Sharpe ratio difference to the MKT strategy

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
US	**0.24**	**0.26**	**0.17**	**0.26**	**0.22**	**0.19**	**0.20**	**0.19**	**0.22**	**0.31**	**0.23**	**0.25**	**0.28**	**0.18**	**0.22**	**0.21**	0.13	**0.18**
	4.28	*4.95*	*2.87*	*4.91*	*3.68*	*3.33*	*3.38*	*2.92*	*3.65*	*5.09*	*3.40*	*3.46*	*4.72*	*2.67*	*3.23*	*2.94*	*1.82*	*2.98*
DM	**0.20**	0.16	0.05	**0.24**	0.08	0.07	**0.23**	0.11	**0.20**	**0.26**	0.11	**0.22**	0.13	0.13	0.24	**0.45**	0.21	**0.23**
	3.03	*2.04*	*0.84*	*3.02*	*0.73*	*0.98*	*3.26*	*0.80*	*2.43*	*2.59*	*0.84*	*2.31*	*1.22*	*0.93*	*2.10*	*3.41*	*1.48*	*2.15*
EM	0.12	0.03	0.01	0.13	0.01	0.08	**0.15**	0.02	**0.14**	0.17	0.02	**0.29**	**0.26**	0.03	**0.31**	**0.24**	−.08	0.16
	1.75	*0.37*	*0.09*	*1.53*	*0.18*	*1.17*	*2.15*	*0.25*	*2.19*	*1.83*	*0.20*	*3.66*	*3.86*	*0.30*	*3.75*	*2.97*	*0.68*	*1.84*

Panel B: Information ratio difference to the ND strategy

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
US	0.30	0.34	–	0.29	0.12	–	0.38	0.22	–	**0.53**	0.23	–	**0.58**	0.26	–	0.46	0.21	–
	1.69	*1.89*	–	*1.84*	*0.73*	–	*1.68*	*1.21*	–	*2.55*	*1.33*	–	*2.76*	*1.40*	–	*1.82*	*0.99*	–
DM	0.44	0.23	–	0.59	0.13	–	0.74	0.15	–	0.56	0.22	–	0.61	0.38	–	**0.96**	0.39	–
	1.56	*0.77*	–	*1.76*	*0.34*	–	*1.86*	*0.35*	–	*1.53*	*0.56*	–	*1.38*	*0.91*	–	*2.55*	*1.03*	–
EM	0.51	0.09	–	0.27	−.21	–	0.48	−.08	–	0.09	−.47	–	0.67	−.25	–	**1.05**	−.21	–
	1.31	*0.24*	–	*0.80*	*0.53*	–	*1.33*	*0.22*	–	*0.25*	*1.16*	–	*1.93*	*0.61*	–	*2.58*	*0.58*	–

Averages:

Panel A	MS	1M	ND
Diff SR	0.23	0.11	0.18
t-value	3.11	1.57	2.39

Panel B	MS	1M
Diff IR	0.53	0.10
t-value	1.73	0.82

Table 7: Multiple Tries Over Time

Monthly alpha in percentage of the Markov switching (MS) strategy (Alpha), the single hypothesis p-value of the robust alpha test of Leippold and Rüegg (2018) (Pval), and the multiple hypothesis adjusted p-value by the method of Romano and Wolf (2016) (Padj) that controls the FWER. In Panel A we show the analysis for a holding period of one month, in Panel B for a holding period of one month including transaction costs, and in Panel C for a monthly holding period for the US, a quarterly holding period for the DM, and a six-months holding period for the EM. In each Panel, we show the results for the six factor sets FF to DMNU from Table 3. The analysis period is from April 2003 to December 2017. We highlight significant multiple hypothesis adjusted alphas below the 5% (10%) significance level in bold (italics).

Panel A: Excluding transaction costs (monthly holding period)

	FF			FFC			CLV			CB			CLVB			DMNU		
	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM
Alpha	**0.28**	**0.27**	0.12	0.16	*0.34*	0.11	**0.44**	**0.42**	0.28	**0.51**	*0.33*	0.25	**0.42**	*0.37*	**0.51**	*0.39*	**0.59**	*0.41*
Pval	0.00	0.00	0.27	0.09	0.02	0.44	0.00	0.00	0.10	0.00	0.02	0.15	0.00	0.01	0.01	0.01	0.00	0.01
Padj	0.04	0.04	0.45	0.33	0.08	0.45	0.02	0.02	0.33	0.00	0.08	0.34	0.02	0.08	0.04	0.07	0.01	0.08

Panel B: Including transaction costs (monthly holding period)

	FF			FFC			CLV			CB			CLVB			DMNU		
	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM
Alpha	0.19	0.08	−.37	0.07	0.16	−.39	0.35	0.20	−.34	**0.41**	0.13	−.35	0.33	0.16	−.07	0.28	0.37	−.24
Pval	0.02	0.33	0.01	0.43	0.24	0.02	0.01	0.09	0.09	0.00	0.33	0.10	0.01	0.24	0.68	0.05	0.02	0.11
Padj	0.31	0.73	0.17	0.73	0.64	0.19	0.10	0.48	0.48	0.02	0.73	0.50	0.10	0.71	0.73	0.33	0.18	0.51

Panel C: Including transaction costs (US monthly, DM quarterly, EM half-yearly holding period)

	FF			FFC			CLV			CB			CLVB			DMNU		
	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM
Alpha	0.19	0.15	0.28	0.07	0.25	0.33	0.35	0.19	0.33	*0.41*	*0.34*	0.33	0.33	0.34	0.39	0.28	0.32	0.35
Pval	0.03	0.12	0.08	0.43	0.03	0.10	0.01	0.13	0.05	0.00	0.01	0.06	0.01	0.01	0.07	0.05	0.05	0.12
Padj	0.32	0.34	0.34	0.43	0.29	0.34	0.15	0.34	0.31	0.05	0.10	0.34	0.15	0.14	0.31	0.34	0.29	0.34

Table A.1: Factor performance

Annualized mean (Ret), annualized standard deviation (Std), Sharpe ratio (SR), as well as minimum (Min) and maximum (Max) monthly return of the factor portfolios. The columns show first the market (MKT) strategy and the top (left) and bottom (right) portfolio of the eight factors from Table 3. We show the returns for the three universes US, DM, and EM from Table 2. The numbers are in percentage.

Panel US: June 1963 to December 2017																	
	MKT	HML		RMW		CMA		WML		LVOL		BETA		SUE		REV	
Ret	5.2	6.5	4.5	6.4	3.7	6.1	4.7	8.0	3.6	5.4	4.7	5.5	4.0	5.9	4.9	6.4	3.5
Std	15.1	14.7	16.5	15.2	16.2	13.8	17.7	17.6	17.8	12.3	23.5	12.2	23.6	15.1	15.4	18.5	16.2
SR	34.5	44.1	27.4	42.2	22.7	44.2	26.3	45.3	20.3	43.6	19.9	44.9	17.0	39.1	31.8	34.6	21.9
Min	22.3	18.9	24.2	22.8	23.0	19.5	24.5	26.0	20.2	17.8	29.7	15.8	27.9	23.3	20.4	23.9	24.4
Max	16.4	22.6	21.1	17.9	14.6	15.0	21.1	20.0	25.0	15.3	23.0	18.9	25.2	18.0	15.7	21.7	17.8
Panel DM: April 1998 to April 2018																	
	MKT	HML		RMW		CMA		WML		LVOL		BETA		SUE		REV	
Ret	4.6	5.7	4.7	6.1	2.6	6.1	4.0	6.8	2.7	5.4	3.0	4.7	2.8	5.0	4.7	4.9	3.0
Std	14.8	18.2	14.5	13.7	16.1	14.3	17.3	15.7	20.6	11.5	23.1	10.7	22.9	15.9	15.5	18.5	15.9
SR	31.3	31.6	32.5	44.6	16.4	42.7	23.3	43.0	12.9	47.4	13.1	44.3	12.1	31.6	30.2	26.3	18.7
Min	17.2	21.9	15.5	14.9	19.3	15.6	21.3	15.8	23.0	13.1	26.3	13.1	26.0	17.9	18.8	26.1	16.0
Max	10.1	19.0	10.8	9.7	11.7	10.6	14.0	17.8	26.8	8.1	19.3	7.5	18.9	10.9	12.1	18.2	17.1
Panel EM: April 1998 to April 2018																	
	MKT	HML		RMW		CMA		WML		LVOL		BETA		SUE		REV	
Ret	7.7	10.6	8.0	10.4	5.0	8.6	7.5	11.4	4.3	8.1	6.2	8.6	5.5	8.9	6.5	8.4	8.5
Std	23.5	29.2	22.7	22.9	26.0	24.7	24.8	24.4	29.1	19.6	30.8	17.4	30.9	24.7	24.2	26.1	23.7
SR	32.5	36.2	35.2	45.4	19.3	34.8	30.1	46.6	14.9	41.5	20.3	49.4	17.7	36.2	26.7	32.1	35.8
Min	29.4	38.3	25.6	27.2	35.4	25.9	30.2	27.6	35.3	29.8	34.2	19.9	38.1	31.0	30.4	30.5	25.8
Max	16.7	28.7	16.5	19.9	20.0	25.2	20.6	16.2	28.2	13.7	23.3	14.3	19.9	21.2	19.0	22.0	20.6

Table B.1: Out-of-sample performance analysis

Excess return (Ret), standard deviation (Std), maximum drawdown (MD), Sharpe ratio (SR), Calmar ratio (CR), as well as the relative return (Rel), tracking error (TE), relative maximum drawdown (RD), information ratio (IR) and relative Calmar ratio to the market (CRr) of the Markov switching (MS), momentum (1M) and naive diversification (ND) strategies. The factor models and universes are the same as in Table 5. We also highlight the return of the market in the header of the Panels. The numbers are annualized and in percentage.

Panel US: MKT with Ret 5.0, SR 32.6, CR 9.0

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
Ret	9.0	9.2	7.2	9.2	8.9	7.6	9.6	8.9	7.1	10.7	9.3	7.6	10.8	8.8	6.9	9.7	8.0	6.6
SR	56.6	58.7	50.0	58.2	54.5	51.8	52.6	51.8	54.4	63.6	55.6	57.5	60.7	50.6	54.2	53.1	45.3	50.4
CR	16.9	20.2	14.6	19.5	18.8	16.5	14.7	17.3	16.2	20.2	20.0	17.5	20.4	17.8	14.9	15.4	12.2	13.4
Rel	3.9	4.1	1.9	4.1	3.8	2.2	4.7	3.8	1.6	5.6	4.2	2.0	5.8	3.7	1.3	4.9	3.0	1.0
IR	62.9	67.2	32.8	71.1	53.2	41.6	63.2	47.6	25.3	82.0	52.1	29.2	76.9	44.7	18.9	63.5	38.6	17.6
CRr	23.6	26.0	5.1	32.6	24.8	10.7	16.8	21.7	4.9	31.6	23.4	5.9	34.5	17.7	3.6	16.1	9.9	3.3

Panel DM: MKT with Ret 8.2, SR 60.5, CR 16.7

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
Ret	11.6	10.6	9.4	12.1	10.5	9.1	12.9	10.9	9.4	11.9	10.8	9.2	12.2	11.3	8.9	14.5	11.3	9.1
SR	80.8	76.9	65.6	84.9	68.3	67.6	83.7	71.0	80.9	86.3	71.7	82.1	73.4	73.8	84.5	105.7	81.2	83.2
CR	25.8	22.9	18.9	27.7	21.1	19.7	27.8	21.8	22.3	28.5	21.7	22.6	25.1	22.7	21.9	40.5	26.9	22.8
Rel	3.4	2.4	1.2	3.9	2.3	0.9	4.9	2.7	0.9	3.6	2.6	0.6	4.2	3.0	0.2	6.2	3.0	0.5
IR	79.5	59.1	35.8	83.4	37.4	24.2	97.2	38.3	23.5	69.4	36.1	13.7	65.1	41.5	3.7	105.9	48.6	9.7
CRr	43.6	30.1	14.3	59.6	21.9	10.4	72.5	24.1	9.3	48.7	14.6	4.0	39.4	23.6	1.2	58.4	33.2	3.6

Panel EM: MKT with Ret 10.2, SR 48.3, CR 16.8

	FF			FFC			CLV			CB			CLVB			DMNU		
	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND	MS	1M	ND
Ret	13.2	11.1	10.7	14.0	11.3	12.2	14.6	11.4	12.0	14.5	10.8	14.2	17.3	11.2	13.5	16.8	8.6	11.0
SR	60.5	51.1	49.1	61.3	49.8	56.0	63.8	50.8	62.6	65.0	50.3	77.3	74.8	51.6	78.8	72.6	40.8	64.1
CR	20.7	17.2	18.6	21.9	17.5	19.6	22.3	17.9	21.2	21.6	17.0	26.7	27.8	18.0	27.1	28.0	13.6	21.2
Rel	3.0	0.9	0.4	4.0	1.2	1.9	4.6	1.1	1.2	4.2	0.3	3.3	7.3	0.8	2.3	6.8	−1.9	−0.2
IR	56.5	14.8	5.9	64.3	17.1	37.8	69.6	13.9	21.9	60.3	4.4	50.9	100.5	9.1	33.7	101.0	−24.8	−3.6
CRr	34.3	5.2	1.7	16.6	7.9	13.4	31.5	5.9	7.9	21.6	1.5	21.3	96.2	4.1	13.5	62.1	−6.8	−1.1

Table C.1: MS strategy's timing ability over different holding periods

Monthly alpha in percentage and the Newey-West t-value with a bandwidth of $4*(T/100)^{(2/9)}$ in italics of the Markov switching strategy (MS) relative to the strategies' underlying factor model. In Panel A we distinguish between the three universes US, DM, and EM from Table 2 and the six factor sets FF to DMNU from Table 3. We highlight in bold significant alphas with the robust alpha test of Leippold and Rüegg (2018) below the 5% significance level. In Panel B we provide the averages among the three universes.

Panel A: Relative to the strategies' underlying factor model

	FF			FFC			CLV			CB			CLVB			DMNU		
	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM
1	**0.33**	**0.28**	0.13	**0.28**	**0.35**	0.11	**0.41**	**0.41**	0.27	**0.51**	**0.34**	0.25	**0.51**	**0.38**	**0.50**	**0.45**	**0.58**	**0.42**
	5.32	*3.85*	*1.26*	*4.66*	*3.22*	*0.89*	*5.24*	*3.99*	*1.85*	*7.52*	*3.09*	*1.61*	*6.44*	*3.01*	*3.44*	*5.65*	*4.62*	*3.27*
2	**0.16**	0.21	**0.35**	**0.16**	0.16	**0.37**	**0.23**	0.25	**0.51**	**0.17**	0.21	0.36	**0.29**	0.23	**0.38**	**0.18**	0.18	**0.34**
	2.35	*2.13*	*3.02*	*2.66*	*2.22*	*3.25*	*2.84*	*1.96*	*4.14*	*2.51*	*1.65*	*2.06*	*3.34*	*2.04*	*2.60*	*3.28*	*1.78*	*2.41*
3	0.05	**0.23**	0.23	0.14	**0.32**	0.25	**0.17**	**0.26**	0.15	**0.22**	**0.40**	**0.62**	**0.21**	**0.41**	**0.50**	**0.19**	**0.39**	**0.44**
	0.77	*2.80*	*1.58*	*1.84*	*3.43*	*1.77*	*2.34*	*2.56*	*1.15*	*2.73*	*4.17*	*4.47*	*2.72*	*3.96*	*3.80*	*3.51*	*3.35*	*4.55*
4	0.10	0.11	**0.26**	0.05	**0.18**	**0.39**	**0.22**	0.16	**0.43**	0.10	0.04	0.38	**0.26**	0.09	0.40	**0.20**	**0.30**	**0.43**
	1.83	*1.31*	*2.36*	*0.84*	*2.44*	*2.46*	*3.24*	*1.97*	*2.62*	*1.50*	*0.30*	*2.24*	*3.13*	*0.83*	*2.22*	*2.61*	*2.43*	*2.68*
6	0.11	**0.19**	**0.36**	0.15	0.10	**0.41**	**0.24**	0.14	**0.39**	0.15	0.27	**0.44**	**0.25**	**0.32**	**0.46**	**0.25**	**0.18**	0.42
	1.95	*2.42*	*2.63*	*2.10*	*1.55*	*2.55*	*3.85*	*1.62*	*2.86*	*1.96*	*2.30*	*2.95*	*3.99*	*3.56*	*2.93*	*3.70*	*2.21*	*2.59*
12	**0.15**	**0.29**	0.13	**0.27**	**0.32**	0.31	**0.20**	**0.43**	0.30	**0.33**	**0.51**	0.29	**0.15**	**0.50**	0.30	0.11	**0.45**	0.42
	2.34	*4.37*	*1.10*	*3.84*	*3.32*	*1.94*	*2.85*	*4.19*	*1.69*	*4.49*	*4.27*	*2.03*	*2.34*	*3.56*	*1.76*	*2.02*	*4.22*	*2.87*

Panel B: Averages across the universes

US						DM						EM					
1	2	3	4	6	12	1	2	3	4	6	12	1	2	3	4	6	12
0.41	0.20	0.16	0.15	0.19	0.20	0.39	0.21	0.34	0.14	0.20	0.42	0.28	0.38	0.37	0.38	0.41	0.29
5.81	*2.83*	*2.32*	*2.19*	*2.92*	*2.98*	*3.63*	*1.96*	*3.38*	*1.54*	*2.28*	*3.99*	*2.05*	*2.91*	*2.89*	*2.43*	*2.75*	*1.90*

Table C.2: MS strategy's Sharpe ratio differences to the market across different holding periods

Annualized Sharpe ratio difference and the Newey-West t-value with a bandwidth of $4 * (T/100)^{(2/9)}$ in italics of the Markov switching strategy (MS) relative to the strategies' underlying factor model. In Panel A we distinguish between the three universes US, DM, and EM from Table 2 and the six factor sets FF to DMNU from Table 3. We highlight in bold significant alphas by the robust Sharpe ratio test of Ledoit and Wolf (2008) below the 5% significance level. In Panel B we provide the averages among the three universes.

Panel A: Sharpe ratio difference compared to the MKT strategy

	FF			FFC			CLV			CB			CLVB			DMNU		
	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM
1	**0.24**	**0.20**	0.12	**0.26**	**0.24**	0.13	**0.20**	**0.23**	**0.15**	**0.31**	**0.26**	0.17	**0.28**	0.13	**0.26**	**0.21**	**0.45**	**0.24**
	4.28	*3.03*	*1.75*	*4.91*	*3.02*	*1.53*	*3.38*	*3.26*	*2.15*	*5.09*	*2.59*	*1.83*	*4.72*	*1.22*	*3.86*	*2.94*	*3.41*	*2.97*
2	**0.14**	0.09	**0.23**	**0.18**	0.07	**0.23**	**0.13**	0.07	**0.24**	0.05	0.13	**0.21**	**0.16**	0.04	**0.21**	−.03	−.04	0.16
	2.89	0.85	3.56	3.27	0.88	2.84	2.41	0.65	3.33	0.91	1.08	2.34	2.85	0.45	2.72	0.66	0.39	2.07
3	0.02	**0.19**	0.12	**0.14**	**0.18**	0.18	0.06	0.06	0.10	0.12	0.20	**0.28**	0.03	0.16	**0.20**	−.03	0.19	**0.24**
	0.38	2.27	1.66	2.36	2.06	1.97	0.97	0.63	1.33	1.95	2.08	3.68	0.50	1.76	2.44	0.63	1.59	3.69
4	**0.13**	0.04	**0.18**	0.09	0.11	**0.16**	**0.11**	0.03	0.15	0.06	−.03	0.14	0.10	−.03	0.16	−.02	0.07	0.25
	2.46	0.57	2.23	1.77	1.84	2.08	2.04	0.29	1.78	1.30	0.27	1.95	1.63	0.30	1.70	0.38	0.69	2.15
6	0.05	0.11	0.13	0.10	0.08	0.16	0.09	0.05	0.13	**0.12**	0.03	0.15	0.07	0.13	**0.20**	**0.11**	0.15	**0.30**
	1.18	1.38	1.47	1.93	0.97	1.89	1.54	0.54	1.77	2.15	0.33	2.02	1.23	1.64	2.51	2.32	1.67	4.68
12	**0.17**	**0.29**	0.07	**0.25**	**0.28**	**0.22**	0.10	**0.34**	0.18	**0.16**	**0.39**	**0.25**	0.05	**0.43**	**0.31**	−.07	**0.41**	**0.30**
	2.78	4.31	1.07	4.05	2.69	2.90	1.64	3.39	1.63	3.14	3.05	3.25	0.98	3.05	3.32	1.36	4.23	3.94

Panel B: Averages across the universes

	US						DM						EM					
	1	2	3	4	6	12	1	2	3	4	6	12	1	2	3	4	6	12
	0.25	0.11	0.06	0.08	0.09	0.11	0.25	0.06	0.16	0.03	0.09	0.36	0.18	0.21	0.19	0.17	0.18	0.22
	4.22	*2.17*	*1.13*	*1.60*	*1.72*	*2.33*	*2.76*	*0.71*	*1.73*	*0.66*	*1.09*	*3.45*	*2.35*	*2.81*	*2.46*	*1.98*	*2.39*	*2.69*

Table C.3: MS strategy's information ratio difference to the naive diversification strategy across different holding periods

Annualized Information ratio difference and the Newey-West t-value with a bandwidth of $4 * (T/100)^{(2/9)}$ in italics of the Markov switching strategy (MS) relative to the naive diversification strategy (ND). In Panel A we distinguish between the three universes US, DM, and EM from Table 2 and the six factor sets FF to DMNU from Table 3. We highlight in bold significant alphas by the robust information ratio test of Ledoit and Wolf (2008) below the 5% significance level. In Panel B we provide the averages among the three universes.

Panel A: Information ratio difference compared to the ND strategy

	FF			FFC			CLV			CB			CLVB			DMNU		
	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM	US	DM	EM
1	0.30	0.44	0.51	0.29	0.59	0.27	0.38	0.74	0.48	**0.53**	0.56	0.09	**0.58**	0.61	0.67	0.46	**0.96**	**1.05**
	1.69	*1.56*	*1.31*	*1.84*	*1.76*	*0.80*	*1.68*	*1.86*	*1.33*	*2.55*	*1.53*	*0.25*	*2.76*	*1.38*	*1.93*	*1.82*	*2.55*	*2.58*
2	0.11	0.07	0.76	0.11	0.30	0.49	0.26	0.35	**1.00**	0.07	0.52	0.37	0.38	0.53	0.69	0.05	0.47	**0.96**
	0.69	*0.20*	*1.98*	*0.64*	*0.93*	*1.29*	*1.08*	*0.77*	*2.60*	*0.27*	*1.22*	*0.93*	*1.66*	*1.18*	*1.80*	*0.18*	*0.76*	*2.15*
3	−.20	0.56	0.26	0.11	0.59	0.19	0.17	0.48	0.13	0.29	**0.84**	0.51	0.24	0.83	0.58	0.16	0.79	**0.86**
	0.91	*1.53*	*0.72*	*0.61*	*1.64*	*0.45*	*0.64*	*1.04*	*0.37*	*1.14*	*2.12*	*1.35*	*0.92*	*1.84*	*1.21*	*0.62*	*1.70*	*2.29*
4	0.08	0.12	0.48	−.06	0.72	0.36	0.29	0.59	0.63	0.18	0.36	0.24	0.37	0.48	0.50	0.17	0.62	**0.96**
	0.46	*0.33*	*1.21*	*0.29*	*2.03*	*0.95*	*1.18*	*1.05*	*1.46*	*0.74*	*0.61*	*0.65*	*1.41*	*0.90*	*1.29*	*0.61*	*1.12*	*2.30*
6	−.07	0.33	0.30	0.10	0.40	0.42	0.29	0.44	0.47	0.31	0.48	0.09	0.35	0.90	0.33	0.41	0.67	**0.96**
	0.30	*1.08*	*0.92*	*0.50*	*1.43*	*1.26*	*1.13*	*0.89*	*1.26*	*1.42*	*1.02*	*0.25*	*1.29*	*1.97*	*0.95*	*1.93*	*1.77*	*2.77*
12	0.18	**1.01**	0.51	0.34	**0.93**	0.50	0.35	**1.22**	0.49	0.47	**1.00**	0.43	0.29	**1.15**	**0.60**	0.06	**1.35**	**1.05**
	0.99	*2.49*	*1.28*	*1.95*	*2.74*	*1.34*	*1.37*	*3.25*	*1.50*	*2.05*	*2.81*	*1.43*	*1.14*	*3.03*	*2.13*	*0.23*	*3.55*	*3.06*

Panel B: Averages across the universes

US						DM						EM					
1	2	3	4	6	12	1	2	3	4	6	12	1	2	3	4	6	12
0.42	0.16	0.13	0.17	0.23	0.28	0.65	0.37	0.68	0.48	0.53	1.11	0.51	0.71	0.42	0.53	0.43	0.60
2.05	*0.75*	*0.81*	*0.78*	*1.10*	*1.29*	*1.77*	*0.84*	*1.65*	*1.01*	*1.36*	*2.98*	*1.36*	*1.79*	*1.06*	*1.31*	*1.23*	*1.79*

III

Appendix

Curriculum Vitae

Personal Data

Name	Roger
Surname	Rueegg
Date of Birth	15. November 1987

Education

Sep 2013 – Feb 2019	PhD Student at Department of Banking & Finance, University of Zurich, Switzerland
Sep 2010 – Sep 2012	Master of Science in Quantitative Finance, University of Zurich & ETH Zurich, Switzerland
Sep 2007 – Oct 2010	Bachelor of Arts in Banking and Finance, University of Zurich, Switzerland
Aug 2000 – Aug 2006	Swiss Matura, Kantonsschule Rychenberg, Winterthur, Switzerland

Work Experience

Mar 2012 – present	Portfolio Manager at Swisscanto Invest, Zürcher Kantonalbank, Zurich, Switzerland
Jan 2012 – Feb 2012	Trainee in the Start-up Finance division, Zürcher Kantonalbank, Zurich, Switzerland
Oct 2012 – Dec 2012	Quantitative Analyst, Systematic Alpha Management LLC, New York, USA
Aug 2011 – Sep 2012	Intern & Trainee in the Asset Management division, Zürcher Kantonalbank, Zurich, Switzerland
Aug 2008 – Jul 2011	Corporate Actions Officer, Credit Suisse AG, Zurich, Switzerland